SOMMAIRE

I- Le modèle probabiliste

Voici les premières phrases d'un manuel [1]: "La théorie des probabilités est une science mathématique étudiant les lois régissant les phénomènes aléatoires. Un phénomène est aléatoire si, reproduit maintes fois, il se déroule chaque fois un peu différemment, de sorte que le résultat de l'expérience change d'une fois à l'autre d'une manière aléatoire, imprévisible."

L'usage même du mot expérience sous-entend que le phénomène aléatoire est observé par le biais d'un critère bien défini, et que le résultat de cette observation peut être décrit sans ambiguïté. L'expérience peut aussi être répétée, et on suppose que chacun des résultats possibles est observé avec une certaine fréquence dont la valeur se stabilise si on répète l'expérience maintes et "maintes fois". C'est cette "loi" que présuppose l'existence d'un modèle probabiliste.

Ce premier chapitre est une rapide présentation du cadre formel des *modèles probabilistes.*

1- Evènements

Etant donnée une expérience aléatoire, on note Ω l'ensemble de tous les résultats possibles de cette expérience.

Un singleton de Ω est appelé *évènement élémentaire.*

Un sous-ensemble A de Ω est appelé un *évènement* . Un évènement A est donc un ensemble constitué de résultats possibles de l'expérience. Si le résultat d'une expérience est dans A, on dit que A est réalisé.

Exemple 1-1 : On détermine le sexe d'un nouveau-né. On posera :

$$\Omega = \{g, f\}$$

Le résultat g signifie que le nouveau-né est un garçon et f que c'est une fille.•

Exemple 1-2 : Sept étudiants doivent passer un oral d'examen. On leur distribue un numéro d'ordre. On pose :

Ω = {tous les alignements des sept lettres a, b, c, d, e, f, g}

Le résultat cfabdeg signifie que l'étudiant c est le premier, a le second, ...

L'ensemble des arrangements qui commencent par cf est un évènement.•

[1] H.Ventsel : Théorie des probabilités. (Ed.MIR, traduction française 1973).

Exemple 1-3 : L'expérience consiste à déterminer la dose d'anésthésique minimale (exprimée en ml) à administrer à un patient pour l'endormir. On choisit :

$$\Omega =]\,0,\, +\infty[$$

L'évènement] 2, 3] est réalisé si la dose minimale à administrer est comprise entre 2 et 3, c'est-à-dire si une quantité supérieure ou égale à 3 suffit à endormir le patient, mais une quantité inférieure à 2 est insuffisante.•

Dans le cadre de la théorie des probabilités, un évènement est généralement défini comme l'ensemble des résultats ayant une propriété donnée. La plupart du temps, l'ensemble A est noté comme la propriété qui le définit. Donnons quelques exemples de telles assimilations :

Ω	:	évènement certain
$\emptyset$	:	évènement impossible
$A \cup B$	:	évènement (A ou B)
$A \cap B$	:	évènement (A et B)
A^c	:	(non A), évènement contraire de A
$A \cap B = \emptyset$	:	les évènements A et B sont incompatibles

Exercice 1-1 : Soit Ω l'ensemble des résultats possibles d'une expérience aléatoire, et soient A, B et C des évènements. Traduire en termes ensemblistes les évènements :

a) les trois évènements A, B et C sont réalisés

b) aucun des évènements A, B ou C n'est réalisé

c) au moins un des évènements est réalisé

d) deux au plus des évènements est réalisé

2- Loi de probabilité, espace de probabilité

On tire une boule dans une urne contenant 2 boules blanches, 1 noire, 4 vertes, 5 rouges, et on regarde sa couleur. Si on répète cette expérience, la fréquence avec laquelle on obtient une boule rouge se stabilise peu à peu sur une valeur, égale ici à 5/12. On dit couramment qu'on a 5 chances sur 12 de tirer une boule rouge. Dans le cadre d'un modèle mathématique de cette expérience aléatoire, on dira que l'évènement "tirer une boule rouge" a la probabilité 5/12. Plus généralement, dans un modèle probabiliste, chaque évènement est pondéré par un nombre compris entre 0 et 1, sa probabilité. Ces probabilités doivent respecter certaines règles de compatibilité, naturelles si on les interprète en termes de "nombre de chances sur 100". L'additivité est la principale de ces règles. Appliquée à un cas particulier dans notre exemple, elle exprime simplement que, puisqu'on a 5 chances sur 12 de tirer une boule rouge et 2 chances sur 12 de tirer une

blanche, on a 5+2 chances sur 12 de tirer une boule soit rouge soit blanche. L'autre règle dit seulement que si on tire une boule, on a 100% de chances de …tirer une boule…

Définition 1-1 : Soit Ω un ensemble. Une *loi de probabilité* P sur Ω est une fonction qui à tout évènement A associe un nombre réel P(A), et qui a les trois propriétés :

a) $0 \leq P(A) \leq 1$,

b) $P(\Omega) = 1$

c) Pour toute famille finie ou dénombrable $(A_n)_{n\in I}$ d'évènements deux à deux disjoints :

$$P(\biguplus_{n\in I} A_n) = \sum_{n\in I} P(A_n) \ .$$

(Ω, P) s'appelle un *espace de probabilité*. •

Exemple 1-4 : On lance un dé et on observe la face du dessus. On posera :

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

et on supposera que le dé est parfaitement équilibré, de sorte que la probabilité de chaque face est la même :

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6} .$$

Remarquons qu'alors, la probabilité de tout évènement est calculable en utilisant la propriété c) de la définition. Par exemple, comme {1, 3, 4} est la réunion des trois ensembles 2 à 2 incompatibles {1}, {3} et {4}, on a :

$$P(\{1, 3, 4\}) = P(\{1\}) + P(\{3\}) + P(\{4\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \ .\bullet$$

Plus généralement, soit Ω un ensemble fini :

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$$

Définir une loi de probabilité P sur Ω revient à se donner n réels positifs ou nuls p_1, p_2, ..., p_n tels que $\sum_{k=1}^{n} p_k = 1$, et à poser, pour tout indice k, $P(\{\omega_k\}) = p_k$. La loi de probabilité sur Ω est alors complètement déterminée car, étant donné un évènement A, P(A) est calculable en additionnant les probabilités p_k de chacun des évènements élémentaires $\{\omega_k\}$ qui composent A.

Il en est de même si Ω est un ensemble dénombrable, les sommes finies sont alors remplacées par les sommes de séries.

Exercice 1-2 : Soit (Ω, P) un espace de probabilité. Répondre aux questions en utilisant la définition 1-1 :

a) Si A est un évènement de probabilité P(A) connue, que vaut $P(A^c)$?

b) Si $A \subset B$, comparer P(A) et P(B).

c) Calculer P(A ou B) en fonction de P(A et B), P(A) et P(B).

d) Montrer que P(A ou B) ≤ P(A)+P(B). Généraliser cette inégalité à un nombre fini d'évènements.

On pourrait aussi démontrer les propriétés suivantes :

Proposition 1-1 : a) Pour toute famille finie ou dénombrable $(A_n)_{n\in I}$ d'évènements :

$$P(\bigcup_{n\in I} A_n) \leq \sum_{n\in I} P(A_n) \ .$$

b) Si $(A_n)_{n\in \mathbb{N}}$ est une suite croissante d'évènements :

$$P(\bigcup_{n\in \mathbb{N}} A_n) = \lim_{n \to +\infty} P(A_n)$$

c) Si $(A_n)_{n\in \mathbb{N}}$ une suite décroissante d'évènements :

$$P(\bigcap_{n\in \mathbb{N}} A_n) = \lim_{n \to +\infty} P(A_n) \bullet$$

3- Le cas où les évènements élémentaires sont équiprobables

Soit (Ω, P) un espace de probabilité correspondant à une expérience aléatoire dont l'ensemble des résultats possibles est fini :

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$$

Supposons que chaque résultat "a autant de chances d'être réalisé qu'un autre", soit, en termes probabilistes, que P est telle que :

$$P(\{\omega_1\}) = P(\{\omega_2\}) = ... = P(\{\omega_n\})$$

Comme la somme de ces n nombres est 1, leur valeur commune est égale à 1/n . Soit maintenant un évènement A. Sa probabilité est :

$$P(A) = \sum_{k \ / \ \omega_k \in A} P(\{\omega_k\}) = \text{card}(A) \ . \ \frac{1}{n} = \frac{\text{card}(A)}{\text{card}(\Omega)}$$

Cette loi de probabilité est souvent appelée *loi uniforme* sur Ω. Calculer des probabilités par une méthode directe dans ce cas revient donc à dénombrer des ensembles.

Exercice 1-3 : Un jeune enfant qui ne sait pas lire prend les 6 jetons d'un jeu de Scrabble qui composaient le mot "CARTON". Il réaligne ces jetons au hasard. Avec quelle probabilité recompose-t-il ce mot ? Même question s'il a pris les 8 jetons qui composaient le mot "INSTITUT".

Exercice 1-4 : 20 sujets sont au programme d'un oral d'examen. Le candidat tire au sort 3 de ces sujets et traite l'un de ces trois. Combien doit-il avoir révisé de sujets pour avoir au moins 9 chances sur 10 de pouvoir traiter un sujet qu'il a révisé ?

Remarque sur le choix du modèle probabiliste

Comme dans tout problème de modélisation, il n'y a pas d'automatisme qui permette d'associer un espace de probabilité à une expérience aléatoire "concrète". Même dans des

cas d'école, il n'y a jamais un seul "bon" choix : reprenons l'exemple de l'urne introduisant le paragraphe 2. Deux modèles peuvent être considérés comme naturels :

- On peut distinguer les 12 boules contenues dans l'urne en posant :

$$\Omega = \{B1, B2, N, V1, V2, V3, V4, R1, R2, R3, R4, R5\}$$

On munit alors Ω de la probabilité uniforme.

- On peut aussi choisir de ne représenter que la couleur de la boule tirée, en posant :

$$\Omega = \{B, N, V, R\}$$

et en définissant P par :

$P(\{B\}) = 2/12$ $P(\{N\}) = 1/12$ $P(\{V\}) = 4/12$ $P(\{R\}) = 5/12$.

Il est clair cependant qu'il est difficile de justifier le deuxième modèle sans faire appel à l'idée d'équiprobabilité des tirages, idée qui par contre est clairement exprimée dans le premier modèle.

Un autre exemple, celui-là célèbre, est du type de celui de l'*aiguille de Buffon* : quelle est la longueur moyenne d'une corde d'un cercle de rayon r, comment représenter le tirage au hasard d'une telle corde ?

Dans des cas concrets de modélisation, les hypothèses sur lesquelles reposent la définition du modèle doivent être clairement énoncées, de telle sorte qu'elles puissent être commentées et éventuellement remises en question, soit directement, soit par leurs implications théoriques, soit par une confrontation avec des données expérimentales.

4- Exercices

Exercice 1-5 : Soit (Ω, P) un espace de probabilité, et soient A et B deux évènements.

Montrer que si $P(A) = P(B) = 0,9$, alors, $P(A \cap B) \geq 0,8$.

Dans le cas général, montrer que $P(A \cap B) \geq P(A) + P(B) - 1$.

Exercice 1-6 : Deux personnes sont tirées au sort dans un groupe de 30 composé de 10 femmes et 20 hommes. Avec quelle probabilité ces deux personnes sont-elles des hommes ? Avec quelle probabilité sont-elles des femmes ?

Exercice 1-7 : Deux amis font partie d'un groupe de n personnes, auxquelles on a distribué au hasard des numéros d'ordre pour constituer une file d'attente.

a) Avec quelle probabilité sont-ils les deux premiers ?

b) Avec quelle probabilité sont-ils distants de r places, c'est-à-dire séparés par r-1 personnes. Représenter ces probabilités par un diagramme en bâtons.

Exercice 1-8 : Un tiroir contient en vrac les 20 chaussettes de 10 paires différentes. On en sort au hasard 4 chaussettes. Avec quelle probabilité obtient-on :

a) 2 paires b) au moins une paire

II- Probabilités conditionnelles

1- Définition

Lançons un dé parfaitement équilibré. Un bon modèle probabiliste en est donné par :

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

muni de la loi de probabilité P uniforme.

Notons A l'évènement "le dé donne au moins 4 points" et B l'évènement "le résultat est impair". Supposons qu'on ne retienne le résultat du lancer que s'il est dans B. Dans cette nouvelle expérience, l'évènement A est réalisé quand on obtient un 5, et c'est avec la *probabilité relative* $\frac{P(\{5\})}{P(\{1, 3, 5\})} = \frac{1/6}{3/6} = 1/3$. Plus généralement la probabilité relative de A sous la condition que B est réalisé est $\frac{P(A \text{ et } B)}{P(B)}$. On l'appelle aussi probabilité de A sachant que B, ou probabilité conditionnelle de A relative à B, etc…

Définition 2-1 : Soit (Ω, P) un espace de probabilité, et soit B un évènement tel que $P(B) \neq 0$. La *probabilité de A sachant que B* est notée $P(A \mid B)$, et est définie par :

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \bullet$$

Exercice 2-1 : a) Soit B un évènement tel que $P(B) \neq 0$. Montrer que l'application qui à A associe $P(A \mid B)$ est une loi de probabilité sur Ω.

b) Donner une propriété de A qui implique $P(A \mid B) = 1$, qui implique $P(A \mid B) = 0$, qui implique $P(A \mid B) = \frac{P(A)}{P(B)}$.

Exercice 2-2 : Un couple a deux enfants. Sous l'une des conditions suivantes :

a) l'aîné est un garçon,

b) l'un des enfants est un garçon,

avec quelle probabilité le couple a-t-il un fils et une fille ?

2- Deux résultats de décomposition

Les deux résultats de ce paragraphe utilisent "à l'envers" la définition 2-1, c'est-à-dire donnent un moyen de calcul de probabilités connaissant des probabilités conditionnelles. Ils sont très utiles dans la pratique.

Exemple 2-1 : Une urne contient deux boules blanches et une boule noire. Une personne tire une boule et la garde, une deuxième personne tire une boule. Avec quelle probabilité les deux boules tirées sont-elles blanches ? On peut répondre à cette question en utilisant la définition 2-1. En effet, notons A l'évènement "la première personne a tiré une boule blanche" et B l'évènement "la deuxième personne a tiré une boule blanche". D'après la définition, P(A et B) = P(B | A) P(A). Mais P(A) est connue, c'est 2/3. P(B | A) est aussi connue : c'est 1/2 car, la première personne ayant tiré une boule blanche, la deuxième personne tire une boule au hasard dans une urne qui contient une boule blanche et une boule noire. Ainsi, P(A et B) vaut (2/3).(1/2) = 1/3 .•

La proposition suivante, parfois appelé "théorème des probabilités composées", généralise ce procédé de calcul :

Proposition 2-1 : Soit (Ω, P) un espace de probabilité, et soient A_1, A_2,…, A_n des évènements. On a :

$P(A_n \text{ et } A_{n-1} \text{ et} \ldots \text{ et } A_1) =$

$= P(A_n \mid A_{n-1} \text{ et} \ldots \text{ et } A_1)\ P(A_{n-1} \mid A_{n-2} \text{ et} \ldots \text{ et } A_1) \ldots P(A_2 \mid A_1)\ P(A_1)$. •

Cet énoncé est constamment utilisé dans le contexte des "chaînes de Markov", qui interviennent naturellement dans les problèmes concrets où A_1, A_2,…, A_n représente une succession (temporelle) d'évènements, la probabilité de réalisation du n-ième évènement A_n étant conditionnée par "le passé" (probabilité sachant que A_1 et … et A_{n-1} ont eu lieu). En voici un exemple simple :

Exercice 2-3: On sait que si le flash d'un appareil photo n'a pas eu panne durant les n premiers déclenchements (n entier positif ou nul), la probabilité pour qu'il fonctionne au (n+1)-ième est égale à p ($0 < p < 1$).

a) Quel est la probabilité pour qu'il n'ait pas de panne au cours des 100 premiers déclenchements ?

b) Sachant qu'il a fonctionné n fois, avec quelle probabilité fontionnera-t-il au moins 100 fois de plus ?

Soient C_1, C_2, …, C_n n évènements deux à deux disjoints et dont la réunion est l'ensemble de tous les résultats possibles Ω. En termes ensemblistes, $\{C_1, C_2, \ldots, C_n\}$ est donc une partition de Ω ; en termes probabilistes, on l'appelle un *système complet d'évènements* . Soit A un évènement. On a bien sûr :

$$A = (A \cap C_1) \cup (A \cap C_2) \cup \ldots \cup (A \cap C_n)$$

et les ensembles $(A \cap C_1)$, $(A \cap C_2)$, …, $(A \cap C_n)$ sont deux à deux disjoints. Ainsi :

$$P(A) = P(A \cap C_1) + P(A \cap C_2) + \ldots + P(A \cap C_n)$$

et en utilisant la définition 2-1, on obtient le résultat :

Proposition 2-2 : Soit (Ω, P) un espace de probabilité, et soit $\{C_1, C_2, \ldots, C_n\}$ un système complet d'évènements. Soit A un évènement. On a :

$$P(A) = P(A \mid C_1)\, P(C_1) + P(A \mid C_2)\, P(C_2) + \ldots + P(A \mid C_n)\, P(C_n) \quad \bullet$$

(Remarquons sans démonstration que ce résultat se généralise à un système complet dénombrable d'évènements.)

Exercice 2-4 : En mars 1994 (enquête sur l'emploi INSEE 1994), la population active en France comprend 44,7% de femmes. Le taux de chômage chez les hommes est 10,8% ; il est chez les femmes 14,3% . On tire au sort une personne parmi les actifs.

a) Avec quelle probabilité est-elle au chômage ?

b) Sachant qu'elle est au chômage, avec quelle probabilité est-ce une femme ?

3- Evènements indépendants

Il est naturel de poser que, du point de vue de leur probabilité de réalisation, deux évènements A et B sont indépendants si le fait de savoir que B est réalisé n'apporte pas d'information sur les chances de réalisation de A, c'est-à-dire si la probabilité de A sachant que B est égale à P(A), et donc si $P(A \cap B) = P(A)\, P(B)$. Posons pour définition plus générale la suivante :

Définition 2-2 : Soit (Ω, P) un espace de probabilité, et soit $(A_i)_{i \in I}$ une famille d'évènements. On dit que ces évènements sont *indépendants dans leur ensemble* si, quelle que soit la partie finie J de I, $P(\bigcap_{j \in J} A_j) = \prod_{j \in J} P(A_j)$. •

Exercice 2-5 : a) Montrer que si A et B sont indépendants, A et B^c, A^c et B, A^c et B^c le sont aussi. Généraliser cette remarque au cas d'une famille finie d'évènements indépendants dans leur ensemble.

b) Deux évènements A et B incompatibles sont-ils indépendants ?

c) Par un diagramme donner un exemple d'évènements A, B, C deux à deux indépendants mais qui ne sont pas indépendants dans leur ensemble.

Remarque : Lançons deux dés, chacun parfaitement équilibré. L'ensemble des résultats possibles est :

$$\Omega = \{ (i, j),\ 1 \le i \le 6,\ 1 \le j \le 6 \} = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$$

Notons A l'évènement "le premier dé donne 4". Comme le premier dé est parfaitement équilibré, la probabilité de A est 1/6. Notons B l'évènement "le deuxième dé donne 6". Comme le deuxième dé est parfaitement équilibré, la probabilité de A est 1/6. De plus, nous pouvons sans difficulté supposer que les évènements A et B sont indépendants. Donc, la probabilité de (A et B), c'est-à-dire de l'évènement élémentaire (4, 6), est égale à (1/6).(1/6) = 1/36, et de même bien sûr pour tout autre couple (i, j). Ce raisonnement confirme le choix de la loi uniforme sur Ω pour représenter l'expérience aléatoire du lancer de deux dés.

Exercice 2-6 : On lance deux dés. Avec quelle probabilité la somme des points obtenus est-elle égale à 11 ? à 10 ?

Plus généralement, considérons une expérience aléatoire dont (Ω, P) est un modèle probabiliste. Si cette expérience est répétée n fois de façon indépendante, on choisira comme ensemble de résultats $\tilde{\Omega} = \Omega^n$, qu'on munira de la *probabilité produit* $\tilde{P}$, c'est-à-dire telle que, quels que soient les sous-ensembles $A_1, A_2,\ldots, A_n$ de Ω :

$$\tilde{P}(A_1 \times A_2,\ldots \times A_n) = P(A_1)\, P(A_2)\ldots P(A_n)\ . \bullet$$

4- Exercices

Exercice 2-7 : Avec quelle probabilité une famille de 3 enfants comporte-t-elle au moins un garçon ?

Exercice 2-8 : Dans un groupe de 20 personnes, quelle est la probabilité pour qu'il n'y ait jamais plus d'un anniversaire par jour ? Et dans un groupe de 50 personnes ? (on fera comme si toutes les années avaient 365 jours).

Exercice 2-9 : Une expérience est conduite pour étudier la mémoire des rats. Un rat est mis devant trois couloirs. Au bout de l'un d'eux se trouve de la nourriture qu'il aime, au bout des deux autres, il reçoit une décharge électrique. Cette expérience élémentaire est répétée jusqu'à ce que le rat trouve le bon couloir. Sous chacune des hypothèses suivantes :

(H1) le rat n'a aucun souvenir des expériences antérieures,

(H2) le rat se souvient de l'expérience immédiatement précédente,

(H3) le rat se souvient des deux expériences précédentes,

avec quelle probabilité la première tentative réussie est-elle la k-ième ? Représenter graphiquement les réponses.

Exercice 2-10 : Pour décider d'un traitement thérapeutique, on utilise un test qui est positif 99 fois sur 100 si une personne est effectivement malade. Mais si une personne n'est pas malade, le test est positif une fois sur 100. On sait par ailleurs que 5 personnes sur 100 ont cette maladie.

a) Si le test d'une personne est positif, avec quelle probabilité cette personne est-elle effectivement malade ?

b) Si le test d'une personne est négatif, avec quelle probabilité cette personne n'est-elle effectivement pas malade ?

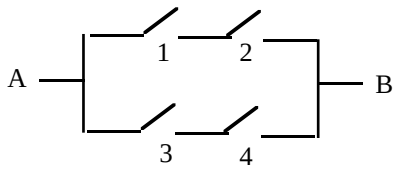
Calculer ces probabilités quand on sait que 5 personnes sur 1000 ont cette maladie.

Exercice 2-11 : La probabilité de fermeture du relai i des circuits décrits ci-dessous est p_i. Tous les relais fonctionnent indépendamment. Dans chacun des cas suivants, quelle est la probabilité pour que le courant passe entre A et B ?

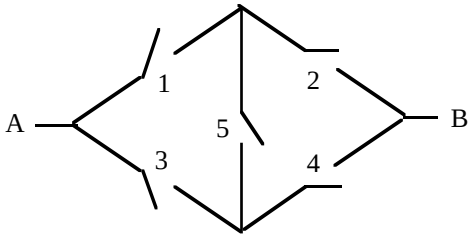
a) A et B sont séparés par n relais reliés en série.

b) A et B sont séparés par n relais reliés en parallèle.

c)



d)



Exercice 2-12 : On transmet un message composé de n symboles binaires '0' ou '1'. Lors de la transmission, chaque symbole est perturbé avec la probabilité p et se transforme alors en symbole opposé. Par précaution, le message est transmis deux fois. Si les deux messages transmis coïncident, l'information est considérée comme correcte.

a) Avec quelle probabilité le i-ième symbole du premier message transmis est-il identique au i-ième symbole du deuxième message transmis ?

b) Avec quelle probabilité les deux messages transmis sont-ils identiques ?

c) Trouver la probabilité pour que, malgré la coïncidence des deux messages, l'information s'avère erronée. (Application numérique : n = 100 p = 0,001).

Exercice 2-13 : Un candidat d'un jeu télévisé américain est face à trois portes. Derrière l'une d'elles se trouve le prix, - une voiture -. Le candidat se place devant la porte de son choix. Le présentateur de l'émission, qui lui sait où se trouve la voiture, ouvre alors l'une des deux autres portes et indique au candidat que la voiture ne s'y trouve pas. Le candidat peut à son tour ouvrir une porte. S'il découvre la voiture, il la gagne.

Un candidat décide d'adopter l'une des trois stratégies suivantes :

a) ouvrir la porte devant laquelle il s'est placé à l'issu de son premier choix,

b) ouvrir l'autre porte,

c) tirer à pile ou face et, s'il obtient pile, ouvrir la porte devant laquelle il s'est placé à l'issu de son premier choix, ouvrir l'autre porte s'il obtient face.

L'une de ces trois stratégies est-elle préférable aux autres ?

III- Variables aléatoires : généralités

1- Définitions

Dans beaucoup de situations, le détail du résultat d'une expérience aléatoire ne nous intéresse pas, mais seulement une valeur numérique fonction de ce résultat. Par exemple, on peut se demander quel est le nombre de pannes d'un ordinateur sur une durée d'un an, sans être intéressé par les dates auxquelles ont lieu ces pannes. Etudions un exemple plus simple :

Exemple 3-1 : On lance deux dés, et on regarde la somme des points obtenus. On choisit pour modèle probabiliste du lancer des deux dés :

$$\Omega = \{ (i, j) , 1 \leq i \leq 6 , 1 \leq j \leq 6 \}$$

muni de la loi de probabilité P uniforme, qui affecte à chaque évènement élémentaire (i, j) la probabilité $P\{(i, j)\} = 1/36$. Avec quelle probabilité la somme des points obtenus est-elle égale, par exemple, à 5 ? C'est la probabilité de l'ensemble des évènements élémentaires (i, j) qui réalisent cette condition.
Introduisons l'application S de Ω dans $\mathbb{R}$, qu'on dira être une *variable aléatoire* , définie par :

$$\forall (i, j) \in \Omega \qquad S(i, j) = i + j$$

La question posée est le calcul de la probabilité de l'évènement $\{ (i, j) \in \Omega \ / \ S(i, j) = 5 \}$, c'est-à-dire de l'évènement { (1, 4), (2, 3), (3, 2), (4, 1) }. On notera cet évènement, de façon simplifiée, { S = 5 }. On trouve :
$P(\{ S = 5 \}) = P(\{ (i, j) \in \Omega \ / \ S(i, j) = 5 \}) = P(\{ (1, 4), (2, 3), (3, 2), (4, 1) \}) = 4/36$.
Remarquons que S prend ses valeurs dans {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12} et que, par conséquent :

$$\sum_{k=2}^{12} P(\{ S = k \}) = P\left(\biguplus_{k \in \{2,\ldots,12\}} \{ S = k \} \right) = P(\Omega) = 1 \ . \ \bullet$$

Abordons maintenant le cas général, dans lequel l'ensemble des valeurs prises par une variable aléatoire n'est pas forcément fini ou dénombrable :

Définition 3-1 : On appelle *variable aléatoire* une application X définie sur un espace de probabilité (Ω, P) et à valeurs réelles.
La *fonction de répartition* F d'une variable aléatoire X est la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie, pour tout réel x, par :

$$F(x) = P(\{ X \leq x \}) \ \bullet$$

Exercice 3-1 : Représenter la fonction de répartition de la variable aléatoire S de l'exemple 3-1.

Exercice 3-2 : Soit X une variable aléatoire, et soit F sa fonction de répartition. Pour a et b réels (a < b), exprimer en fonction de F :

$$P(X > a), \quad P(a < X \leq b),$$

$P(X < a)$ (utiliser la proposition 1-1-b), $P(X \geq a)$, $P(X = a)$, $P(a \leq X < b)$, …

Cet exercice montre que la connaissance de la fonction de répartition F d'une variable aléatoire X permet de calculer, pour n'importe quel intervalle I de $\mathbb{R}$, la probabilité $P(\{ X \in I \})$. On peut démontrer qu'elle permet aussi, - en principe tout du moins -, de calculer la probabilité $P(\{ X \in B \})$ pour n'importe quel sous-ensemble B de $\mathbb{R}$. On dit en résumé que la fonction de répartition de X détermine la *loi* ou la *loi de probabilité* de X. (Le vocabulaire est justifié par le fait que l'application qui à un sous-ensemble B de $\mathbb{R}$ associe $P(\{ X \in B \})$ est une loi de probabilité sur $\mathbb{R}$).

On peut montrer sans difficulté que, si F est la fonction de répartition d'une variable aléatoire :

a) F est croissante,

b) F est continue à droite en tout point,

c) $\lim_{x \to -\infty} F(x) = 0$ $\qquad$ $\lim_{x \to +\infty} F(x) = 1.$

et inversement, mais la démonstration n'est pas élémentaire, qu'une fonction F de $\mathbb{R}$ dans $\mathbb{R}$ qui vérifie les propriétés a), b) et c) est la fonction de répartition d'une variable aléatoire.

Exercice 3-3: Soit X une variable aléatoire. On suppose que sa fonction de répartition F est donnée par :

$$F(x) = 0 \quad \text{si} \quad x < 0$$
$$= \frac{1}{3} + \frac{2}{3}(1 - e^{-x}) \quad \text{si} \quad x \geq 0$$

a) Dessiner le graphe de F.

b) Calculer :

$P(X > -2)$, $P(X \in]1-1/n, 1])$, $P(X = 1)$, $P(X \in] -1/n, 0])$, $P (X = 0)$

2- Variables aléatoires discrètes, variables aléatoires à densité

- Une variable aléatoire X qui prend ses valeurs dans un sous-ensemble fini ou dénombrable $\{ x_i , i \in I \}$ de $\mathbb{R}$ est dite *discrète*. Notons :

$$p_i = P(X=x_i).$$

Les p_i sont des réels de [0, 1] et tels que $\sum_{i \in I} p_i = 1$.

La donnée des p_i définit la *loi de la variable aléatoire* X, puisque pour tout sous-ensemble A de $\mathbb{R}$:

$$P(X \in A) = \sum_{i \,/\, x_i \in A} p_i \ .$$

La fonction de répartition F de X s'exprime, pour tout a réel, par :

$$F(a) = \sum_{i \,/\, x_i \leq a} p_i$$

Nous avons vu sur un exemple (exercice 3-1) que, tout du moins quand il n'y a qu'un nombre fini de x_i par intervalle borné, F est constante par morceaux, et que ses discontinuités sont situées aux points d'abscisse x_i, la hauteur du saut correspondant étant p_i.

- On dit qu'une variable aléatoire X est *à densité* s'il existe une fonction f de $\mathbb{R}$ dans $\mathbb{R}$, positive ou nulle, et telle que, pour tout sous-ensemble B de $\mathbb{R}$:

$$P(X \in B) = \int_B f(x)\, dx \ .$$

On appelle cette fonction f la *fonction de densité de probabilité* de la variable aléatoire X.

Exercice 3-4 : On fait tourner une aiguille autour d'un axe et on repère la position sur laquelle elle s'arrête par un angle Θ de $[0, 2\pi[$.
a) Quelles valeurs proposer pour $P(0 \leq \Theta < \pi)$, $P(\pi \leq \Theta < 2\pi)$, $P(\pi/2 \leq \Theta < 3\pi/2)$? Et pour $P(\Theta \in I)$ lorsque I est un sous-intervalle de $[0, 2\pi[$?
b) Peut-on proposer une fonction f qui soit la densité de la loi de Θ ?

Remarquons que si X est une variable aléatoire à densité, la densité f vérifie nécessairement :

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1.$$

De plus, quels que soient a et b (a < b) :

$$P(X = a) = 0$$

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

La fonction de répartition F de X est donnée par :

$$F(a) = \int_{-\infty}^{a} f(x)\, dx$$

et elle est continue.

Exercice 3-5 : Soit X une variable aléatoire à densité f définie par :

$$f(x) = c\, x \quad \text{si} \quad 1 \leq x \leq 4$$
$$= 0 \quad \text{sinon}$$

a) Calculer la valeur de c.
b) Que vaut P($1 \le X \le 2$) ?
c) Calculer et représenter graphiquement la fonction de répartition de X.

- Concluons ce paragraphe en signalant qu'une variable aléatoire peut n'être ni discrète, ni à densité, mais *mixte* :

Exercice 3-6 : Reprendre l'exemple de l'exercice 3-3, et montrer qu'on peut écrire :

$$P(X \in B) = \int_B f(x)\,dx \quad \text{si} \quad 0 \notin B$$

$$= \frac{1}{3} + \int_B f(x)\,dx \quad \text{si} \quad 0 \in B ,$$

où f est une fonction à déterminer.

3- Couples de variables aléatoires

Soit (Ω, P) un espace de probabilité, et soient X et Y deux variables aléatoires définies sur cet espace. Le couple (X, Y) définit ce que l'on peut appeler une *variable aléatoire à valeurs dans* $\mathbb{R}^2$: à tout ω de Ω, il associe en effet le vecteur ($X(\omega)$, $Y(\omega)$).

La loi de (X, Y), souvent appelée *loi conjointe* de (X, Y), est déterminée par la donnée, pour tout sous-ensemble C de $\mathbb{R}^2$, de la probabilité P({ $(X, Y) \in C$ }).
On montre que la loi conjointe de (X, Y) est déterminée dès qu'on connaît $P(X \in A \text{ et } Y \in B)$ pour tout couple (A, B) de sous-ensembles de $\mathbb{R}$. On montre aussi qu'il suffit pour cela de connaître la *fonction de répartition* F du couple (X, Y) qui est définie par :

$$\forall (x, y) \in \mathbb{R}^2 \qquad F(x, y) = P(\{ X \le x \text{ et } Y \le y \}) .$$

Remarquons que si la loi conjointe de (X, Y) est connue, on en déduit les lois de X et de Y, appelées dans ce contexte *lois marginales*. En effet, pour tout sous-ensemble A de $\mathbb{R}$:

$$\{ X \in A \} = \{ X \in A \text{ et } Y \in \mathbb{R} \} = \{ (X, Y) \in A \times \mathbb{R} \} ,$$

et on tire :

$$P (X \in A) = P ((X, Y) \in A \times \mathbb{R}) .$$

Exercice 3-7 : Soient (X, Y) un couple de variables aléatoires dont la loi est telle que, si i et j sont deux entiers tels que $0 \le i \le 2$ et $-i \le j \le i$, $P\{ (X, Y) = (i, j) \} = \frac{1}{9}$.

a) Représenter graphiquement les valeurs prises par le couple (X, Y).
b) Quelle sont les lois marginales de X et Y ?

4- Variables aléatoires indépendantes

Définition 3-2 : Soient X et Y deux variables aléatoires définies sur un espace de probabilité (Ω, P). On dit qu'elles sont indépendantes si pour tout couple (A, B) de sous-ensembles de $\mathbb{R}$, les évènements $\{ X \in A \}$ et $\{ Y \in B \}$ sont indépendants, c'est-à-dire si :

$$P (X \in A \text{ et } Y \in B) = P(X \in A) \; P(Y \in B) \bullet$$

Exercice 3-8 : a) Soient (X, Y) un couple de variables aléatoires de loi donnée par :

$$P\{ (X, Y) = (-1, 0) \} = P\{ (X, Y) = (1, 0) \} =$$
$$= P\{ (X, Y) = (0, -1) \} = P\{ (X, Y) = (0, 1) \} = \frac{1}{4} .$$

X et Y sont-elles indépendantes ?

b) même question avec les données de l'exercice 3-7.

Exercice 3-9 : Supposons X et Y discrètes, et plus précisément que X prend ses valeurs dans le sous-ensemble fini ou dénombrable $\{ x_i , i \in I \}$ de $\mathbb{R}$, et que Y prend ses valeurs dans le sous-ensemble fini ou dénombrable $\{ y_j , j \in J \}$ de $\mathbb{R}$. Montrer que X et Y sont indépendantes si et seulement si pour tout couple (i, j) de $I \times J$:

$$P(X=x_i \text{ et } Y=y_j) = P(X=x_i) \; P(Y=y_j).$$

Dans le cas général, on montre la proposition :

Proposition 3-1 : Soient X et Y deux variables aléatoires définies sur un espace de probabilité (Ω, P), de fonctions de répartitions F_X et F_Y. X et Y sont indépendantes si et seulement si, pour tout couple (x, y) de réels :

$$P(X \leq x \text{ et } Y \leq y) = F_X(x) \; F_Y(y) \bullet$$

Le résultat suivant est utile :

Proposition 3-2 : Soient X et Y deux variables aléatoires définies sur un espace de probabilité (Ω, P). Soient ϕ et ψ deux applications de $\mathbb{R}$ dans $\mathbb{R}$. Si X et Y sont indépendantes, alors, $\phi(X)$ et $\psi(Y)$ sont des variables aléatoires indépendantes. $\bullet$

Enonçons enfin une extension de la définition :

Considérons une famille $(X_i)_{i \in I}$ de variables aléatoires définies sur un espace de probabilité (Ω, P).

On dit que c'est une *famille de variables aléatoires indépendantes* si pour toute famille $(A_i)_{i\in I}$ de sous-ensembles de $\mathbb{R}$, les évènements $\{X_i\in A_i\}$ $(i\in I)$ sont indépendants dans leur ensemble, autrement dit si pour tout sous-ensemble fini J de I :

$$P(\ \forall j\in J\ \ X_j\in A_j\ ) = \prod_{j\in J} P(X_j\in A_j).$$

On démontre que si $(X_i)_{i\in I}$ est une famille de variables aléatoires indépendantes, et si J et K sont deux parties finies et disjointes de I décrites par :

$$J =\{j_1, \ldots, j_r\} \qquad\qquad K =\{k_1, \ldots, k_s\}\ ,$$

si ϕ une fonction de $\mathbb{R}^r$ dans $\mathbb{R}$ et ψ une fonction de $\mathbb{R}^s$ dans $\mathbb{R}$, alors, $\phi(X_{j_1},\ldots, X_{j_r})$ et $\psi(X_{k_1},\ldots, X_{k_s})$ sont indépendantes. Et on peut généraliser ce résultat à plusieurs parties finies de I deux à deux disjointes.

5- Exercices

Exercice 3-10 : On équipe un local souterrain de 5 ampoules électriques. On suppose que les durées de vie de ces ampoules sont des variables aléatoires indépendantes, et de même densité f donnée par :

$$f(x) = \frac{200}{x^2} \quad \text{si} \quad x > 200$$
$$= 0 \quad \text{sinon}\ .$$

On contrôle l'état des ampoules après 300 heures d'utilisation. Avec quelle probabilité deux (exactement) des ampoules sont-elles hors d'usage.

Exercice 3-11 : Une boîte contient 5 transistors, dont on sait que 3 sont défectueux. On teste l'un après l'autre les transistors et on les met de côté, jusqu'à avoir trouvé les défectueux. On note N_1 le nombre de tests effectués pour trouver le premier transistor défectueux, et N_2 le nombre de tests complémentaires effectués pour trouver le deuxième. Décrire la loi conjointe de N_1 et N_2.

Exercice 3-12 : Soient $X_1,\ldots, X_n$ des variables aléatoires indépendantes et suivant toutes la loi uniforme sur [0, 1]. On pose :

$$M = \max (X_1,\ldots, X_n)$$

a) Quelle est la fonction de répartition de M ? Quelle est la densité de la loi de M ?
b) Mêmes questions avec $\min (X_1,\ldots, X_n)$.

IV- Caractéristiques numériques des variables aléatoires

1- Espérance

Soit X une variable aléatoire sur un espace de probabilité (Ω, P). L'*espérance* E(X) de X est la valeur moyenne des valeurs prises par X, pondérées par leur probabilité de réalisation. Les mathématiciens disposent d'une théorie, la *théorie de la mesure*, dans laquelle l'intégrale $\int_{\Omega} X(\omega)\, dP(\omega)$ a un sens. Ils définissent E(X) par cette intégrale. Si Ω est fini ou dénombrable, cette intégrale est simplement la somme $\sum_{\omega \in \Omega} X(\omega)\, P(\{\omega\})$, mais le cas général est plus complexe. Ici, nous nous restreignons aux deux cas particuliers des variables aléatoires discrètes ou à densité, et nous utiliserons comme définition de l'espérance les caractérisations suivantes :

- Si X est discrète et prend ses valeurs dans un sous-ensemble fini ou dénombrable $\{ x_i , i \in I \}$:

$$E(X) = \sum_{i \in I} x_i\, P(X = x_i)$$

- Si X est à densité f :

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

Exercice 4-1 : Quelle est l'espérance de la variable aléatoire qui représente le nombre de points obtenus en lançant un dé ?

Exercice 4-2 : Quelle est l'espérance de la variable aléatoire de l'exercice 3-4 ?

Exercice 4-3 : Dans chacun des deux cas suivants, calculer E(X), décrire la loi de X^2 et calculer $E(X^2)$:

a) P(X = -2) = 0,1 P(X = 1) = 0,6 P(X = 2) = 0,3

b) X à densité f définie par :

$$f(x) = 1/2 \quad \text{si} \quad -1 \leq x \leq 1$$
$$= 0 \quad \text{sinon .}$$

Pour calculer l'espérance X^2 ou plus généralement d'une fonction $\phi(X)$ de X, on peut éviter la détermination de la loi de $\phi(X)$ en utilisant le résultat suivant :

Proposition 4-1 : Soit X une variable aléatoire et soit ϕ une fonction de $\mathbb{R}$ dans $\mathbb{R}$.

- Si X est discrète et prend ses valeurs dans un sous-ensemble fini ou dénombrable $\{ x_i , i \in I \}$:

$$E(\phi(X)) = \sum_{i \in I} \phi(x_i)\, P(X = x_i)$$

- Si X est à densité f :

$$E(\phi(X)) = \int_{-\infty}^{+\infty} \phi(x)\, f(x)\, dx \quad \bullet$$

Exercice 4-4 : Reprendre les exemples de l'exercice 4-3 et calculer $E(X^2)$ en utilisant la proposition 4-1.

L'énoncé suivant sera très utilisé par la suite :

Proposition 4-2 : Soient X et Y deux variables aléatoires sur un espace de probabilité (Ω, P), et soient a et b deux réels. Alors :

$$E(aX+b) = aE(X) + b$$
$$E(X+Y) = E(X) + E(Y) \quad \bullet$$

Exercice 4-5 : Montrer la deuxième égalité de cette proposition dans le cas où les lois de X et Y sont discrètes.

Exercice 4-6 : On lance deux dés, et on note S la variable aléatoire qui représente la somme des points obtenus. Quelle est l'espérance de S ?

2- Variance, covariance

Exemple 4-2 : Considérons les quatre variables aléatoires :

$X_1 = 0$, c'est-à-dire la variable "aléatoire" constante et nulle,

X_2 de loi uniforme sur [-1, 1]

X_3 de loi uniforme sur [-100, +100]

X_4 telle que $P(T=-3000) = 1/2 \quad P(T=2000) = P(T=4000) = 1/4$

Elles ont toutes quatre pour espérance 0, mais leurs lois sont clairement différentes. Une caractéristique qui les distingue est l'étalement, la dispersion, des valeurs qu'elles prennent autour de leur valeur moyenne $E(X_i) = 0$. Une façon de mesurer cette dispersion est de regarder la valeur moyenne de la distance entre X_i et $E(X_i)$. Pour des raisons pratiques, on préfère choisir la valeur moyenne du carré de la distance entre X_i et $E(X_i)$, qu'on appelle la variance. $\bullet$

Définition 4-1 : Soit X une variable aléatoire sur un espace de probabilité (Ω, P). La *variance* $\nu(X)$ de X est :

$$\nu(X) = E[\ (X-E(X))^2\ ]$$

L'*écart-type* $\sigma(X)$ de X est :

$$\sigma(X) = \sqrt{\nu(X)}\ \bullet$$

(Remarquons que si l'unité de mesure dans laquelle X est exprimé est, par exemple, le mètre, $\nu(X)$ est en m^2 et $\sigma(X)$ en mètre).

De l'égalité :

$$[X-E(X)]^2 = X^2 - 2\ E(X)\ X + [E(X)]^2$$

on déduit :

$$\nu(X) = E[\ X^2 - 2\ E(X)\ X + (E(X))^2\ ] = E(\ X^2\ ) -2\ E(X)\ E(X) + [E(X)]^2$$

et finalement :

$$\nu(X) = \ E(\ X^2\ ) - [E(X)]^2$$

Cette égalité est souvent utile dans le calcul effectif de variances.

Exercice 4-7 : Calculer les variances des variables aléatoires X_i de l'exemple 4-2.

Exercice 4-8 : On lance un dé, et on note X la variable aléatoire qui représente le nombre de points obtenus. Quelle est la variance de X ?

Proposition 4-3 : Soit X une variable aléatoire.

a) La variance de X est nulle si et seulement si il existe un réel c tel que P(X=c) = 1. On dit alors que X est *presque sûrement* constante.

b) Soient a et b deux réels. Alors :

$$\nu(aX+b) = a^2\ \nu(X) \qquad\qquad \sigma(aX+b) = a\ \sigma(X)\ \bullet$$

Dans le cas où X n'est pas presque sûrement constante, on remarquera que la variable aléatoire $\dfrac{X - E(X)}{\sigma(X)}$ a son espérance nulle, et un écart-type égal à 1. Elle est ce qu'on appelle la *variable aléatoire centrée réduite* associée à X. Le passage de l'une des variables à l'autre se fait tout simplement par un changement d'origine et d'unité dans l'ensemble des valeurs prises par X.

L'expression de la variance d'une variable aléatoire n'est manifestement pas linéaire. De fait, si X et Y sont deux variables aléatoires sur (Ω, P), en général, la variance de la somme X+Y n'est pas égale à la somme des variances de X et de Y :

Exemple 4-3 : Soit par exemple X une variable aléatoire de variance non nulle, - c'est-à-dire qui n'est pas presque sûrement constante -. On a :

$$\nu(X + (-X)) = \nu(0) = 0 \quad \text{et} \quad \nu(X) + \nu(-X) = 2\nu(X) \neq 0 \ . \bullet$$

Calculons dans le cas général $\nu(X+Y)$. Comme :

$$(X+Y) - E(X+Y) = (X - E(X)) + (Y - E(Y)) \ ,$$

on a :

$$[\ (X+Y) - E(X+Y)\ ]^2 = [X - E(X)]^2 + [Y - E(Y)]^2 + 2\ [X - E(X)]\ [Y - E(Y)]$$

d'où :

$$\nu(X+Y) = \nu(X) + \nu(Y) + 2\ E[\ (X - E(X))\ (Y - E(Y))\ ]$$

Introduisons la définition de la *covariance* de X et Y :

$$\mathrm{cov}(X, Y) = E[\ (X - E(X))\ \ (Y - E(Y))\ ]$$

Ce terme n'est en général pas nul. Cependant :

Proposition 4-4 : Soient X et Y deux variables aléatoires sur (Ω, P). Si X et Y sont indépendantes, alors :

$$\mathrm{cov}(X, Y) = 0$$
$$\nu(X+Y) = \nu(X) + \nu(Y) \ . \bullet$$

Pour montrer ce résultat, on commence par montrer que si X et Y sont indépendantes, $E(XY) = E(X)\ E(Y)$, et conlut en remarquant que sous cette même hypothèse, les variables aléatoires $(X - E(X))$ et $(Y - E(Y))$ sont indépendantes, ou encore en montrant l'égalité $\mathrm{cov}(X, Y) = E(XY) - E(X)\ E(Y)$.

Exercice 4-9 : Démontrer la proposition dans le cas où les lois de X et Y sont discrètes.

Exercice 4-10 : On lance deux dés, et on note S la variable aléatoire qui représente la somme des points obtenus. Quelle est la variance de S ?

Une caractéristique souvent utilisée en statistiques est un coefficient appelé *coefficient de corrélation* de deux variables aléatoires X et Y. C'est par définition, - et si ni X ni Y n'est presque sûrement constante - :

$$\rho\ (X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma(X)\ \sigma(Y)} \ .$$

Remarquons que c'est un coefficient sans dimension.

On peut montrer par des méthodes classiques en analyse que :

$$-1 \leq \rho\ (X, Y) \leq 1$$

$\rho\ (X, Y) = 1$ si et seulement si il existe $a > 0$ et b réel tel que $Y = aX + b$

$\rho(X, Y) = -1$ si et seulement si il existe $a < 0$ et b réel tel que $Y = aX + b$.

Méfions-nous cependant : le fait que le coefficient de corrélation de X et Y est nul ne signifie pas du tout que X et Y sont indépendantes (…qu'il n'y a pas de corrélation entre X et Y…). Prenons par exemple une variable X de loi symétrique par rapport à 0 (par exemple de loi uniforme sur [-1, 1]), et posons $Y = X^2$. La loi de $XY = X^3$ est aussi symétrique par rapport à 0. Ainsi, $E(XY) = 0 = E(X)\,E(Y)$, et donc $\rho(X, Y) = 0$. Pourtant, X et Y ne sont pas (du tout) indépendantes, puisqu'au contraire, la donnée de la valeur prise par X détermine complètement la valeur prise par Y.

3- Exercices

Exercice 4-11 : Calculer l'espérance et la variance de la variable aléatoire M de l'exercice 3-12.

Exercice 4-12 : Les transistors fournis par une usine sont défectueux dans la proportion p. On teste un transistor après l'autre jusqu'à en obtenir un bon. On note N le nombre de tests effectués. Quelle est la loi de N ? Calculer l'espérance de N.

Exercice 4-13 : Une machine est constituée de n sous-unités identiques. Elle fonctionne si toutes ses sous-unités fonctionnent. Le procédé de construction des sous-unités est tel qu'elles sont défectueuses dans la proportion p, et indépendamment les unes des autres. Pour construire une machine sans défaut, deux procédés sont envisagés :

a) On construit une sous-unité, on la teste, si elle est bonne, on la monte, sinon, on la jette, etc… On continue jusqu'à avoir monté les n sous-unités de la machine. On suppose pour simplifier qu'il n'y a pas de problème de montage. La machine ainsi construite est donc bonne.

b) On construit et monte sans les tester n sous-unités, et on teste la machine ainsi constituée. Si elle ne marche pas, on la jette, et on recommence jusqu'à obtenir une bonne machine.

On note : c_u le coût de construction d'une sous-unité,
t_u le coût du test d'une sous-unité,
t_m le coût du test d'une machine,

et on suppose pour simplifier que le coût d'assemblage des unités est nul.

1) On note C le coût de construction d'une bonne machine. Calculer l'espérance de C dans les deux cas a) et b).

2) On suppose $t_u = t_m = \frac{c_u}{2}$, et n = 10 (puis n = 100). Suivant la valeur de p, quel est le procédé de fabrication qui est préférable ?

V- Variables aléatoires usuelles

Voici une liste de définitions et propriétés de quelques lois connues. On pourra trouver beaucoup d'autres lois classiques dans la "littérature" : les lois géométrique (exercice 4-12), hypergéométrique, multinomiale, gamma, etc…, et nous en introduirons d'autres dans la partie "statistiques" de ce cours.

1- Loi de Bernoulli $\mathcal{B}(p)$

Soit A un évènement de probabilité p. Introduisons la variable aléatoire X telle que :

$X(\omega) = 1$ si $\omega \in A$,

$= 0$ sinon .

On dit que X suit la loi de Bernoulli de paramètre p.

Plus généralement, soit p dans [0, 1]. X suit la loi de Bernoulli $\mathcal{B}(p)$ si :

$P(X=1) = p$ et $P(X=0) = 1 - p$.

$E(X) =$ $\nu(X) =$

2- Loi binomiale $\mathcal{B}(n, p)$

Exercice 5-1 : On lance 4 fois un dé. On note X le nombre de fois où on obtient 6.

a) Pour k = 0, 1, 2, 3, 4, calculer P(X = k).

b) On note X_i la variable de Bernoulli qui vaut 1 si on tire un 6 au i-ième lancer, 0 si on ne tire pas 6 à ce lancer. Ecrire X en fonction des X_i , et en déduire la valeur de E(X) et de $\nu(X)$.

Plus généralement, la loi binomiale $\mathcal{B}(n, p)$ est la loi d'une somme X de n variables aléatoires indépendantes suivant chacune la même loi de Bernoulli $\mathcal{B}(p)$. C'est aussi le nombre de réalisations d'un évènement A lors de l'exécution de n expériences aléatoires indépendantes, le résultat de chacune réalisant A avec la probabilité p. On a :

$P(X = k) =$ (k = 0, 1, …, n)

$E(X) =$ $\nu(X) =$

3- Loi uniforme

La loi uniforme sur intervalle [a, b] de $\mathbb{R}$ est la loi de densité f :

$f(x) = \frac{1}{b-a}$ si $a \le x \le b$

$= 0$ sinon .

$$E(X) = \frac{a+b}{2} \qquad \nu(X) = \frac{(b-a)^2}{12}$$

Exercice 5-2 : Soit X une variable aléatoire de loi uniforme sur [0, 1].
a) Calculer directement E(X) et ν(X).
b) On pose Y = a + (b-a) X . Que valent E(Y) et ν(Y) ? Quelle est la loi de Y ? Qu'en conclut-on ?

4- Loi exponentielle

Soit λ un paramètre strictement positif. La loi exponentielle de paramètre λ est la loi de densité f définie par :

$$f(x) = \lambda e^{-\lambda x} \quad \text{si } x \geq 0$$
$$= 0 \quad \text{sinon .}$$

Si X suit cette loi :

$$E(X) = \frac{1}{\lambda} \qquad \nu(X) = \frac{1}{\lambda^2}$$

On peut remarquer aussi que pour tout t positif ou nul :

$$P(X \geq t+x \mid X \geq t) = P(X \geq x \mid X \geq 0)$$

Cette égalité permet d'interpréter X comme la durée de vie d'un appareil "sans vieillissement" ; en effet, étant donné un instant t, si l'appareil n'est pas tombé en panne auparavant (si $X \geq t$), la probabilité pour qu'il marche encore sans problème durant la période de temps x ($X \geq t+x$) ne dépend pas d'instant t. (Nous avons étudié dans l'exercice 2-3 une situation analogue mais dans le cas discret).
Nous ne détaillerons pas davantage les conditions d'utilisation de cette loi, ni de la loi de Poisson définie dans le paragraphe suivant : ce serait plutôt du ressort d'un cours sur les processus stochastiques.

5- Loi de Poisson $\mathcal{P}(\lambda)$

Soit λ un paramètre strictement positif. On dit que X suit la loi de Poisson $\mathcal{P}(\lambda)$ si X prend ses valeurs dans $\mathbb{N}$ et :

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad (k = 0, 1, 2, \ldots)$$

Cette loi décrit le nombre d'évènements intervenant dans un intervalle de temps de longueur 1, lorsque les laps de temps séparant deux évènements sont indépendants et de même loi exponentielle de paramètre λ.
On a :

$$E(X) = \lambda \qquad \nu(X) = \lambda$$

6- Loi normale $\mathcal{N}(\mu, \sigma)$

La loi normale centrée réduite $\mathcal{N}(0,1)$ est la loi de densité f définie par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} .$$

Si X suit $\mathcal{N}(0,1)$:

$$E(X) = 0 \qquad \nu(X) = 1$$

Exercice 5-3 : On dit qu'une variable aléatoire X suit la loi normale $\mathcal{N}(\mu, \sigma)$ si $\frac{X-\mu}{\sigma}$ suit la loi normale centrée réduite. Si X suit la loi $\mathcal{N}(\mu, \sigma)$, quelles sont les espérance, variance, densité de la loi de X ?

On peut montrer que la primitive de la fonction $e^{-\frac{x^2}{2}}$ ne peut pas s'exprimer à l'aide de fonctions usuelles. La fonction de répartition d'une variable normale se calcule donc point par point et numériquement (voir la table en fin de polycopié) :

Exercice 5-4 : a) Soit X une variable aléatoire de loi $\mathcal{N}(0,1)$. Que valent :

$$P(X \leq -1) \qquad P(-1 < X < 2) \ ?$$

b) Soit X une variable aléatoire de loi $\mathcal{N}(1,4)$. Que vaut $P(X > 5)$?

On montre le résultat important suivant :

Proposition 5-1 : Soient X et Y deux variables aléatoires indépendantes et de lois normales. Alors X+Y suit une loi normale. Plus précisément, si X suit la loi $\mathcal{N}(\mu_1, \sigma_1)$ et Y suit $\mathcal{N}(\mu_2, \sigma_2)$, alors X+Y suit la loi $\mathcal{N}(\quad , \quad)$.•

7- Exercices

Exercice 5-5 : On a constaté que les disquettes produites dans une usine sont défectueuses avec une probablité 0,01 indépendamment les unes des autres. L'usine conditionne ses disquettes par boîtes de 10, et offre à l'acheteur le remboursement d'une boîte dès qu'au moins deux des 10 disquettes sont défectueuses. Dans quelle proportion les boîtes sont-elles renvoyées ? Si quelqu'un achète 3 boîtes, avec quelle probabilité renvoit-il exactement une boîte ? au moins une boîte ?

Exercice 5-6 : On a constaté que le nombre N de clients visitant par jour le magasin d'un tapissier suit une loi de Poisson de paramètre 4, et que chaque client passe une commande

avec la probabilité 0,1 . On note C le nombre de commandes passées par jour. Quelle est la loi de C ? Enoncer un résultat plus général.

Exercice 5-7 : Le diamètre (exprimé en cm.) des tomates livrées à une usine d'emballage américaine suit une loi normale $\mathcal{N}$ (7, σ), où σ est inconnu. Un tri automatique rejette toutes les tomates dont le diamètre n'est pas compris entre 6cm et 8 cm.

a) On constate que 10% des tomates livrées sont rejetées par ce procédé de tri. Calculer l'écart type σ.

b) Le directeur veut réduire à 5% le pourcentage de tomates rejetées lors du tri. Ne pouvant agir sur les livraisons, il installe un système de tri qui rejette les tomates de diamètre inférieur à (7-s) ou supérieures à (7+s). Calculer s.

Exercice 5-8 : On a constaté qu'en absence d'épidémie, la variable aléatoire qui représente le poids d'un poulet de 81 jours pris au hasard dans un élevage des Landes suit une loi normale $\mathcal{N}$ (1,8 , 0,2), et que les poulets se développent indépendamment. On note X la moyenne arithmétique des poids de 100 poulets pris au hasard. Avec quelle probabilité a-t-on (1,79 < X < 1,81) ? Même question en remplaçant 100 par 1000 poulets.

VI- Somme d'un grand nombre de variables aléatoires indépendantes

1- L'inégalité de Tchebychev

Soit X une variable aléatoire d'espérance E(X) et de variance $\nu(X)$, et soit a un réel positif.

Notons Y la variable aléatoire définie par :

$$Y(\omega) = a^2 \quad \text{si} \quad |X(\omega) - E(X)| \geq a$$
$$= 0 \quad \text{sinon .}$$

On a bien sûr :

$$Y \leq [X - E(X)]^2 ,$$

et donc :

$$E(Y) \leq E[(X - E(X))^2] .$$

De plus :

$$E(Y) = a^2 P\{ |X - E(X)| \geq a \}$$

On a ainsi obtenu l'*inégalité de Tchebychev* :

$$P\{ |X - E(X)| \geq a \} \leq \frac{\nu(X)}{a^2} .$$

Cette inégalité n'a bien sûr pas d'intérêt lorsque les probabilités $P(|X - E(X)| \geq a)$ peuvent être calculées explicitement et exprimées simplement. Elle est par contre utile dans le cas contraire, à condition bien sûr de connaître l'espérance et la variance de X. Comme la définition même de $\nu(X)$, l'inégalité de Tchebychev met en évidence l'intérêt de la variance comme mesure de l'étalement des valeurs prises par X autour de la valeur moyenne E(X).

Exercice 6-1 : On lance n fois un dé, et on note M la moyenne arithmétique des points obtenus.

a) Calculer E(M) et $\nu(M)$.

b) Combien de fois suffit-il de lancer un dé pour que, avec une probabilité supérieure à 0,9 , la moyenne arithmétique des points obtenus soit comprise entre 3,4 et 3,6 ?

2- Loi des grands nombres

Considérons n variables aléatoires $X_1,\ldots, X_n$, toutes suivant la même loi d'espérance μ et d'écart-type σ, et intéressons-nous à la moyenne arithmétique M de ces variables aléatoires :

$$M = \frac{X_1 + \ldots + X_n}{n} .$$

Un calcul simple donne l'espérance et la variance de M :

$$E(M) = \mu \qquad\qquad \nu(M) = \frac{\sigma^2}{n}$$

Si a est strictement positif, on a, en vertu de l'inégalité de Tchebychev :

$$P\{ \mid M - \mu \mid \geq a \} \leq \frac{\sigma^2}{na^2} .$$

On en déduit le théorème connu sous le nom de *loi faible des grands nombres* :

Théorème 6-1 : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et de même loi, d'espérance μ et d'écart-type σ. Alors, pour tout a strictement positif :

$$\lim_{n \to +\infty} P\{ \mid \frac{X_1 + \ldots + X_n}{n} - \mu \mid > a \} = 0.$$

On dit que la suite $(\frac{X_1 + \ldots + X_n}{n})_{n\in\mathbb{N}}$ *converge en probabilité* vers μ.•

Exemple 6-1 : Exécutons une suite d'expériences aléatoires indépendantes, le résultat de chacune réalisant un évènement A avec la probabilité p. Pour décrire cette expérience, introduisons les variables aléatoires $X_1,\ldots, X_n,\ldots$ définies par :

$X_i = 1$ si le résultat de la i-ième expérience est dans A,

$= 0$ si le résultat de la i-ième expérience n'est pas dans A.

Ces variables aléatoires sont indépendantes, et suivent toutes la même loi de Bernoulli $\mathfrak{B}(p)$, dont l'espérance est p. La variable aléatoire $\frac{X_1 + \ldots + X_n}{n}$ représente la fréquence de réalisation de A au cours des n premières expériences. On conclut de la loi des grands nombres que, lorsque n est grand, cette fréquence est, dans le sens précisé par l'énoncé, proche de p. Or, p n'est autre que la probabilité de réalisation de A lors d'une expérience. Ainsi, on a construit la théorie mathématique des probabilités en partant de la définition intuitive de la probabilité d'un évènement A comme fréquence de réalisation de A sur un grand nombre d'expériences, et, par une déduction interne au cadre formel mathématique, on démontre cette même propriété.•

3- Théorème central-limite

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et de même loi, d'espérance μ et d'écart-type σ. On a alors :

$$E(X_1 + \ldots + X_n) = n\,\mu , \qquad \nu(X_1 + \ldots + X_n) = n\,\sigma^2 ,$$

et la variable aléatoire centrée réduite associée à la somme $X_1 + \ldots + X_n$ est :

$$Y_n = \frac{X_1 + \ldots + X_n - n\,\mu}{\sigma\,\sqrt{n}}$$

Remarquons que lorsque la loi des X_n est la loi normale $\mathcal{N}(\mu, \sigma)$, la loi de Y_n est la loi normale centrée réduite $\mathcal{N}(0, 1)$. Le théorème central-limite affirme que dans le cas général, la loi $\mathcal{N}(0, 1)$ est une bonne approximation de la loi de Y_n , sous réserve que n soit assez grand. Plus précisément :

Théorème 6-2 : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et de même loi, d'espérance μ et d'écart-type σ. Alors, quel que soit le réel x :

$$\lim_{n \to +\infty} P\left\{ \frac{X_1 + \ldots + X_n - n\,\mu}{\sigma\,\sqrt{n}} \leq x \right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt \ .$$

On dit que $\dfrac{X_1 + \ldots + X_n - n\,\mu}{\sigma\,\sqrt{n}}$ *converge en loi* vers la loi normale $\mathcal{N}(0, 1)$.•

Dans la pratique, cet énoncé rigoureux est interprété assez librement : on considèrera souvent par exemple que sous les hypothèses du théorème, pour des valeurs de n assez grandes, on peut remplacer dans les calculs la loi de $X_1 + \ldots + X_n$ par un loi normale (et donc par la loi normale $\mathcal{N}(n\,\mu, \sigma\,\sqrt{n})$).

Exemple 6-2 : On repère la position d'un point matériel sur une droite par son abscisse. On se donne aussi un réel positif h "petit". Partons de l'abscisse 0. Lançons une pièce. Si on obtient pile, avançons le point d'une distance h, si on obtient face, reculons le point d'une distance h. Et recommençons… C'est un exemple de *marche aléatoire* .
Notons X_n la variable aléatoire définie par :

$X_n = 1$ si on obtient pile au n-ième lancer de la pièce
$= -1$ si on obtient face au n-ième lancer de la pièce.

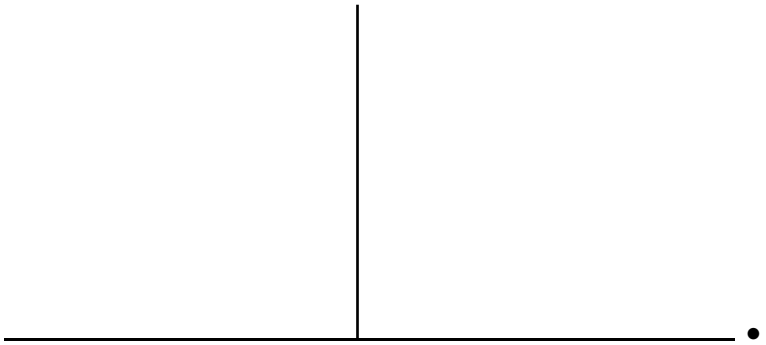
Remarquons que :

$$E(X_n) = 0 \qquad \sigma(X_n) = 1$$

A l'issu du n-ième lancer, le point est situé à l'abscisse $U_n = h\,(X_1 + \ldots + X_n)$. On a :

$$E(U_n) = 0 \qquad \nu(U_n) = nh^2$$

Du théorème central-limite, on déduit que, pour des valeurs de n assez grandes, la loi de U_n est approchée par la loi normale $\mathcal{N}(0, \sqrt{nh^2})$. On peut représenter graphiquement l'évolution de la densité de probabilité de présence du point :



4- Exercices

Exercice 6-2 : Résoudre l'exercice 6-1 en utilisant le théorème central-limite, et comparer les résultats obtenus.

Exercice 6-3 : 60 personnes veulent retirer de l'argent au guichet d'une poste. La somme moyenne demandée par chaque personne est de 400F, avec un écart type de 200F. Les sommes demandées par chaque personne sont indépendantes (et de même loi).
Combien d'argent doit avoir le guichet à sa disposition pour que, avec une probabilité supérieure à 0.95, les 60 personnes retirent la somme qu'elles souhaitent ?

Exercice 6-4 : Une cafétéria d'entreprise fournit chaque jour n repas, et propose chaque jour 2 plats du jour. Le cuisinier a remarqué que lorsqu'il propose saucisse-lentilles et poisson pané-riz, chaque client souhaite le plat de saucisses avec la probabilité $p = 0,6$ et le plat de poisson avec la probabilité 1-p , et que les choix des clients sont indépendants.
Pour tenter de satisfaire sa clientèle, il prépare $np+ s$ plats de saucisses, et $n(1-p)+ s$ plats de riz.
On supposera successivement $n = 100$ et $n = 1000$.
Quelle est la valeur minimale de s telle que, avec une probabilité supérieure à 0,95 , tous les clients aient le plat qu'ils souhaitent ?
Pour cette valeur, quelle est le pourcentage de plats non consommés comparé aux plats préparés ?

Exercice 6-5 : Le nombre de visiteurs potentiels de la Foire de Bordeaux est v=100000. Les visiteurs viennent indépendamment les uns des autres et avec la probabilité p ($0 < p < 1$). On note Y le nombre de personnes qui visitent la foire.
a) Trouver la loi de Y. Quelle sont l'espérance, la variance de Y ?

b) Soit x le prix d'entrée ($x \geq 0$) et R la recette correspondante. Quelle est l'espérance de R ? En supposant p et x reliés par la relation $p = e^{-cx}$, où c est une constante positive, trouver le prix d'entrée qui maximise E(R). Quelle est alors la valeur de E(R) ?
c) Déterminer le nombre maximal n tel que, avec une probabilité supérieure ou égale à 0.8, il y aura au moins n visiteurs.

VII- Echantillonnage

1- Description des données statistiques sur un caractère

On considère ici une *population* , c'est-à-dire un ensemble d'individus. On s'intéresse à un *caractère* particulier des individus de cette population, qu'on suppose, pour chaque individu, quantifiable par un nombre réel. On suppose qu'on a mesuré expérimentalement la valeur du caractère de n individus et qu'on a trouvé les nombres $x_1, \ldots, x_n$.

Exemple 7-1 : La population est l'ensemble des câbles fabriqués dans une usine donnée, le caractère est la charge de rupture d'un câble. On a mesuré la charge de rupture de 12 de ces câbles et obtenu la liste :

1440 1410 1520 1470 1430 1490 1455 1445 1472 1455 1470 1430 •

Exemple 7-2 : La population est l'ensemble des jeux de pile ou face effectués avec une pièce de monnaie donnée, le caractère est égal à 1 si on obtient face et 0 si on obtient pile. On a lancé la pièce 10 fois et obtenu la liste :

0 0 1 0 0 1 0 1 1 0 •

Nous rappelons dans ce paragraphe les outils les plus courants de description des propriétés des listes de résultats $x_1, \ldots, x_n$ obtenues dans ce contexte expérimental.

On peut représenter l'ensemble de ces nombres graphiquement par :

- la *fonction de répartition empirique* : l'ordonnée du point d'abscisse a est égale à $\frac{\text{card}\{j \,/\, x_j \leq a\}}{n}$.

- le *diagramme en bâtons des fréquences* : la hauteur du bâton d'abscisse a est égale à $\text{card}\{j \,/\, x_j = a\}$). Cette représentation n'a d'intérêt que s'il y a des répétitions dans la liste $x_1, \ldots, x_n$.

- un *histogramme des fréquences* : la surface du rectangle de base l'intervalle borné I est égale à $\frac{\text{card}\{j \,/\, x_j \in I\}}{n}$. Un tel histogramme dépend de la façon dont on découpe en intervalles l'ensemble $\mathbb{R}$ des valeurs du caractère.

On peut aussi en calculer des *tendances centrales* :

- la *moyenne* arithmétique $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ,$$

- la *médiane* m : si les x_j sont rénumérotés de telle sorte que $x_{(1)} \leq \ldots \leq x_{(n)}$,

$$m = x_{(k)} \quad \text{si} \quad n = 2k-1$$
$$= \frac{1}{2}\,(x_{(k)} + x_{(k+1)}) \quad \text{si} \quad n = 2k$$

(les *quartiles* , les *déciles* , et plus généralement les *s-quantiles* sont définis de façon analogue en répartissant les $x_{(i)}$ en 4, 10 ou s groupes, au lieu de 2 pour la médiane).

- le *mode* : la valeur a (ou l'une des valeurs) qui maximise card{j / x_j = a}.

On peut en décrire la dispersion par :

- le *rang* : différence entre la plus grande et la plus petite valeur des x_i,

- l'*écart entre certains quantiles* : par exemple, différence entre le troisième et premier quartile,

- l'*écart moyen* : $\dfrac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n}$,

- la *variance empirique* :

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

etc…

Exercice 7-1: Pour l'exemple 7-1, représenter un histogramme de fréquences. Calculer la moyenne $\bar{x}$, la médiane, la variance empirique s^2.

Exercice 7-2 : Pour l'exemple 7-2, calculer la moyenne $\bar{x}$, la variance empirique s^2.

Exercice 7-3 : Proposer une formule récursive pour le calcul conjoint des moyenne et variance empiriques.

On remarque que pour faire ces descriptions, on a au plus besoin d'outils informatiques simples, dans le cas où la liste de données est trop longue pour être traitée "à la main". La situation est différente si on s'intéresse à plusieurs caractères simultanément car les représentations graphiques brutes des données expérimentales sont en générales inexploitables car illisibles (on n'y repère aucun ordre, aucune structure). L'objet de l'*analyse des données* est de proposer des méthodes pour représenter et exploiter de telles données statistiques multivariées.

2- Echantillons aléatoires, statistiques, estimateurs

Reprenons l'exemple 7-1. S'il ne veut pas faire faillite, le fabricant des câbles ne peut pas mesurer la charge de rupture de tous les câbles qu'il fabrique, puisque cette mesure les détruit. La valeur moyenne de la charge de rupture qu'il a calculée en testant 12 câbles reflète-t-elle bien la valeur moyenne de la charge de rupture de l'ensemble des câbles ? La fonction de répartition empirique obtenue est-elle une bonne approximation de celle qu'on obtiendrait après le test de tous les câbles ou de 1200 de ces câbles ? L'objet de la théorie des statistiques est de répondre à des questions de ce type, c'est-à-dire d'estimer la pertinence de la généralisation des caractéristiques de l'échantillon expérimental à la population toute entière.
La démarche choisie est celle de la modélisation probabiliste.
On assimile le caractère numérique, dont $x_1, \ldots, x_n$ est un échantillon observé, à une variable aléatoire X dont la loi est inconnue, ou dont le type est connu mais certains des paramètres sont inconnus. Par exemple, il se peut qu'on sache, pour des raisons théoriques ou en conséquence d'expériences antérieures, que la charge de rupture d'un câble suit une loi normale $\mathcal{N}(\mu, \sigma)$ de paramètres μ et σ inconnus ; mais on peut aussi n'avoir aucune idée a priori sur le type de sa loi.
On représente l'expérience de l'échantillonage par n variables aléatoires $X_1, \ldots, X_n$, indépendantes et de même loi que X, et on considère que la liste $(x_1, \ldots, x_n)$ est un résultat possible de cet expérience, c'est-à-dire une valeur particulière prise par le vecteur aléatoire $(X_1, \ldots, X_n)$.
Nous allons dans ce cours voir comment l'échantillon expérimental peut être utilisé pour estimer la loi de X ou certaines de ses caractéristiques, et donner des moyens de mesurer la validité de ces estimations.

Précisons le cadre de modélisation :

Définition 7-1 : Si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes qui suivent toutes la même loi , on dit qu'elles constituent un *échantillon aléatoire* .•

Définition 7-2 : Soit $X_1, \ldots, X_n$ un échantillon aléatoire. Une *statistique* est une variable aléatoire de la forme $\phi(X_1, \ldots, X_n)$, où ϕ est une fonction déterministe de $\mathbb{R}^n$ dans $\mathbb{R}$. Sa valeur ne dépend que des valeurs prises par $(X_1, \ldots, X_n)$ et non de paramètres de la loi des X_i.•

Exemple 7-3 : Un mois avant un référendum, on sonde 1000 personnes inscrites sur les listes électorales sur leur intention d'aller voter. Les 1000 personnes ont été tirées au hasard (avec remise) dans la population des inscrits. On obtient 650 intentions favorables. Si on tire une personne au hasard et l'interroge, on peut représenter sa réponse par une variable aléatoire X qui vaut 1 si elle a l'intention d'aller voter, et 0 sinon. La loi de X est

une loi de Bernoulli de paramètre p, inconnu, égal à la proportion dans la population des inscrits des personnes ayant l'intention d'aller voter.
On peut représenter l'expérience du sondage en introduisant 1000 variables aléatoires X_1, …, X_{1000} , indépendantes car les 1000 personnes ont été tirées au hasard (avec remise), et qui suivent la loi $\mathfrak{B}(p)$. X_1, …, X_{1000} est un échantillon aléatoire de loi $\mathfrak{B}(p)$. Le sondage a donné des valeurs expérimentales de cet échantillon, x_1, …, x_{1000} , telles que $x_1 + \ldots + x_{1000} = 650$. Une statistique usuelle est $\frac{X_1 + \ldots + X_{1000}}{1000}$. Elle est bien indépendante de l'inconnue p. Sa valeur expérimentale est 0,65 : c'est la proportion parmi les inscrits sondés de personnes ayant l'intention d'aller voter. On étudiera dans la suite du cours dans quelle mesure cette valeur 0,65 peut être considérée comme une approximation de p.•

Dans l'exemple précédent, la statistique $\frac{X_1 + \ldots + X_{1000}}{1000}$ est utilisée pour estimer le paramètre p de la loi $\mathfrak{B}(p)$ de l'échantillon X_1, …, X_{1000} . On dira que c'est un estimateur de ce paramètre p.

Plus généralement, considérons un échantillon aléatoire X_1, …, X_n dont la loi dépend d'un paramètre θ réel (ou vectoriel) inconnu et qu'on veut estimer.

Un *estimateur* du paramètre θ est tout simplement une statistique dont la valeur expérimentale est utilisée comme *estimation* de θ. Un estimateur peut être de plus ou moins bonne qualité, suivant la fiabilité de l'estimation de θ qu'il fournit. Les propriétés qu'on va définir maintenant permettent de cerner la qualité d'un estimateur.

Considérons un estimateur $\Theta_n = \phi_n (X_1, \ldots, X_n)$ de θ, où ϕ_n est une fonction déterministe de $\mathbb{R}^n$ dans $\mathbb{R}$ (ou dans l'espace vectoriel où se trouve θ).

- On dira qu'il est *convergent* si Θ_n converge en probabilité vers θ.
Une traduction de cette condition est qu'il suffit de choisir n assez grand pour que la loi de Θ_n soit aussi resserrée que l'on veut autour de la valeur θ.

- On peut définir l'efficacité de l'estimateur par une mesure du resserrement de la loi de Θ_n autour de θ, qu'on appelle l'*erreur totale* :

$$E((\Theta_n - \theta)^2).$$

Plus cette erreur est petite, - ou converge vite vers 0 - , plus l'efficacité de l'estimateur est grande.

- Par un calcul simple, on montre :

$$E((\Theta_n - \theta)^2) = \nu(\Theta_n) + [E(\Theta_n) - \theta]^2$$

On appelle $(E(\Theta_n) - \theta)$ le *biais* de l'estimateur. Un estimateur est dit *sans biais* si son biais est nul. L'efficacité d'un estimateur sans biais est d'autant plus grande que sa variance est petite, - ou converge vite vers 0 -.

- Par une démonstration semblable à celle de l'inégalité de Tchebychev, on montre aussi que si $E(\Theta_n)$ tend vers θ et $\nu(\Theta_n)$ tend vers 0 quand n tend vers l'infini, Θ_n est un estimateur convergent de θ.

3- Estimateurs les plus usuels

Dans ce paragraphe, $X_1, \ldots, X_n$ désigne un échantillon aléatoire dont la loi a pour fonction de répartition F, μ désigne l'espérance des X_i et σ^2 leur variance.

a) Moyenne de l' échantillon

Définition 7-3 : La *moyenne de l'échantillon* $X_1, \ldots, X_n$ est la variable aléatoire $\bar{X}$:

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \quad \bullet$$

Nous avons déjà vu (inégalité de Tchebychev et loi des grands nombres) que :

$$E(\bar{X}) = \mu \qquad \nu(\bar{X}) = \frac{\sigma^2}{n}$$

La statistique $\bar{X}$ est donc un estimateur sans biais et convergent de l'espérance μ de la loi de X.

b) Variance de l'échantillon

Définition 7-4 : La *variance de l'échantillon* $X_1, \ldots, X_n$ est la variable aléatoire S^2 définie par :

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

L'écart-type de l'échantillon est la variable aléatoire $S = \sqrt{S^2}$.•

Exercice 7-4 : a) Montrer que :

$$(n-1)\, S^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n (\bar{X} - \mu)^2$$

(On a aussi : $(n-1)\,S^2 = \sum_{i=1}^{n} X_i^2 - n\,\bar{X}^2$

Cette formule peut être utile pour les calculs "à la main", mais est numériquement peu fiable. Utiliser l'algorithme de l'exercice 7-3 est préférable pour les calculs sur machine).
b) En déduire que S^2 est un estimateur sans biais de la variance σ^2 de l'échantillon, puis, en utilisant la loi des grands nombres, que cet estimateur est convergent.

Exercice 7-5 : Si l'espérance μ des X_i est connue, montrer que $\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$ est un estimateur sans biais et convergent de la variance σ^2 des X_i .

c) Fonction de répartition de l'échantillon

Définition 7-5 : La *fonction de répartition* F_n de l'échantillon $X_1, \ldots, X_n$ est définie, pour tout x dans $\mathbb{R}$, par :

$$F_n(x) = \frac{\text{card}\ \{\ i\ /\ X_i \leq x\ \}}{n}\ \bullet$$

Remarquons que pour x fixé, $F_n(x)$ est une variable aléatoire réelle. F_n est donc une fonction aléatoire de $\mathbb{R}$ dans $\mathbb{R}$.

Exercice 7-6 : Soit x un réel. Quelle est la loi de $nF_n(x)$? Que vaut $E(F_n(x))$? Montrer que $F_n(x)$ est un estimateur sans biais et convergent de $F(x)$, où F est la fonction de répartition des X_i . (indication : introduire les variables aléatoires I_i définies par :

$I_i(\omega) = 1$ si $X_i(\omega) \leq x$
$= 0$ sinon.)

On pourrait de même définir un histogramme de l'échantillon, etc…

4- Un exemple de comparaison de l'efficacité de deux estimateurs

Supposons ici que la loi de l'échantillon $X_1, \ldots, X_n$ est la loi uniforme sur $[0, 2\theta]$, où θ est le paramètre à estimer.

- Premier estimateur :

Rappelons que si X suit la loi uniforme sur $[0, 2\theta]$, $E(X) = \theta$. Ainsi, $\bar{X}$ est un estimateur sans biais et convergent de θ. Sa variance est :

$$v(\bar{X}) = \frac{v(X)}{n} = \frac{\theta^2}{3n}\ .$$

- Deuxième estimateur :

Considérons l'estimateur :

$$\Theta_n = \frac{\max(X_1, \ldots, X_n)}{2} .$$

La densité d de sa loi (voir l'exercice 3-12) est donnée par :

$$d(x) = \frac{n\, x^{n-1}}{\theta^n} \quad \text{si} \quad 0 \le x \le \theta ,$$

$$= 0 \quad \text{sinon} ,$$

et donc :

$$E(\Theta_n) = \frac{n}{n+1}\ \theta \qquad \nu(\Theta_n) = \frac{n}{(n+2)(n+1)^2}\ \theta^2$$

Θ_n est donc un estimateur convergent de θ.

En posant $\Xi_n = \frac{n+1}{n}\ \Theta_n$, on obtient un estimateur sans biais et convergent de θ, et qui vérifie :

$$\nu(\Xi_n) = \frac{1}{n(n+2)}\ \theta^2$$

On constate que, quel que soit n, il est plus efficace que l'estimateur $\bar{X}$.

Ainsi, dans le cas des lois uniformes sur $[0, 2\theta]$, il est plus efficace, pour estimer l'espérance θ de la loi, d'utiliser un autre estimateur que l'estimateur usuel $\bar{X}$. Mais pour des lois d'un autre type, ce n'est pas forcément le cas. De fait, une méthode de construction d'estimateurs, connue sous le nom de *méthode du maximum de vraisemblance* , produit, pour chaque type de loi et de dépendance par rapport au paramètre, un estimateur de ce paramètre qui est (presque) toujours le plus efficace des estimateurs. On pourra trouver un exposé de cette méthode dans la plupart des manuels de statistiques.

5- Statistiques issues d'une loi normale

a) Lois issues de la loi normale

Définition 7-6 : Soient $Z_1, \ldots, Z_n$ n variables aléatoires indépendantes, qui suivent toutes la loi normale $\mathcal{N}(0, 1)$. La loi du khi-2 à n degrés de liberté, notée χ^2_n, est par définition la loi de $Z_1^2 + \ldots + Z_n^2$.•

Exercice 7-7 : Soit X une variable aléatoire qui suit la loi χ^2_n. Calculer son espérance.

Exercice 7-8 : a) Soit X une variable aléatoire qui suit la loi χ^2_6. Que vaut $P(3 \le X \le 9)$?
b) Soient X et Y deux variables aléatoires indépendantes, X suivant la loi χ^2_3, Y suivant la loi χ^2_6. Que vaut $P(X+Y \ge 10)$?

Définition 7-7 : Soient Z et X deux variables aléatoires indépendantes, Z suivant la loi normale $\mathcal{N}(0, 1)$, et X suivant la loi χ^2_n. La loi de Student à n degrés de liberté, notée t_n, est par définition la loi de $\frac{Z}{\sqrt{X/n}}$.•

On montre qu'une variable aléatoire de Student a une densité symétrique par rapport à 0. Son espérance est donc nulle.

Exercice 7-9 : a) Soit X une variable aléatoire qui suit la loi de Student t_{12}. Que vaut $P(X \leq 1,4)$?
b) Expliquer pourquoi, pour $n \geq 30$, les valeurs données dans la table de la loi de Student sont celles que donnerait l'usage de la table de la loi normale $\mathcal{N}(0, 1)$?

Définition 7-8 : Soient X et Y deux variables aléatoires indépendantes, X suivant la loi χ^2_n, Y suivant la loi χ^2_m. La loi de Fischer à n et m degrés de liberté, notée $F_{n\,m}$, est par définition la loi de $\frac{(\frac{X}{n})}{(\frac{Y}{m})}$.•

b) Moyenne et variance d'un échantillon de loi normale

Proposition 7-1 : Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, $\bar{X}$ et S^2 désignent les moyenne et variance de cet échantillon. Alors :

a) $\bar{X}$ et S^2 sont indépendantes

b) $\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma}$ suit la loi $\mathcal{N}(0 ; 1)$

c) $\frac{(n-1)\,S^2}{\sigma^2}$ suit la loi χ^2_{n-1}

d) $\frac{\sqrt{n}\,(\bar{X} - \mu)}{S}$ suit la loi de Student t_{n-1} •

Exercice 7-10 : On admet les parties a) et c) de l'énoncé. Montrer les autres assertions.

VIII- Tests d'hypothèses sur les valeurs des paramètres d'une variable aléatoire

Dans ce chapitre, $X_1, \ldots, X_n$ désigne un échantillon aléatoire d'une loi qui dépend d'un paramètre réel inconnu θ.

Considérons l'hypothèse sur le paramètre θ :

$$(H) : \theta = \theta_0$$

où θ_0 est un valeur explicite.

On veut construire un test qui utilise un échantillon expérimental $(x_1, \ldots, x_n)$ pour éprouver cette hypothèse. On procède de la façon suivante.

On choisit un estimateur Θ du paramètre θ. On note θ_e sa valeur expérimentale.

On se donne une *variable aléatoire discriminante* de la forme $D = \delta(X_1, \ldots, X_n, \theta_0)$, où δ est une fonction déterministe à valeurs réelles. On fait en sorte que sa valeur expérimentale $d_e = \delta(x_1, \ldots, x_n, \theta_0)$ permette de comparer les valeurs θ_0 et θ_e, en en reflétant par exemple la distance. On se donne aussi une *zone de rejet* R ($R \subset \mathbb{R}$), et on décide :

- de rejeter l'hypothèse (H) si $d_e \in R$
- de considérer que l'expérience ne contredit pas (H) sinon

En prenant cette décision, on court un risque de se tromper.

Si on est dans la deuxième situation, la formulation de la conclusion est tellement molle qu'on ne court pas grand risque de se tromper. (Ce qui ne veut pas dire pour autant que considérer qu'une expérience ne contredit pas (H) est toujours anodin. Supposons par exemple que (H) signifie "Le fonctionnement de la centrale nucléaire de Blaye est normal"...)

Si on est dans la première des situations, il se peut que le paramètre θ soit vraiment égal à θ_0, et que le fait que d_e soit dans la zone de rejet R soit un fait de hasard. Si c'est le cas, le test nous fait rejeter (H) à tort, et plus précisément, si θ_0 est la vraie valeur de θ, en utilisant ce test, on se trompera dans la décision (environ) $100\,\alpha$ fois sur 100, où :

$$\alpha = P\{ D \in R \}.$$

Ce nombre α , ou le pourcentage $(100\,\alpha)\%$, s'appelle le *niveau de risque du test* . Pour pouvoir calculer ce risque, il faut donc choisir la fonction discriminante D de telle sorte que sa loi soit connue lorsque (H) est vraie.

Par la suite, pour construire un test de (H), on fixera en général dès le départ le niveau de risque α, et on définira la zone de rejet R_α de sorte que $P\{ D \in R_\alpha \} = \alpha$.

1- Valeur de l'espérance d'une variable normale de variance connue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où σ est connu et μ inconnu. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) \ : \ \mu = \mu_0$$

Construisons le test.

On utilise l'estimateur $\overline{X}$ du paramètre μ. On note $\overline{x}$ sa valeur expérimentale.
Fixons à α $(0 < \alpha < 1)$ le niveau de risque du test. (Comme α mesure un risque de se tromper, on choisit α "petit", par exemple $\alpha = 0{,}05$ ou $0{,}1$).

Si l'hypothèse (H) est vraie, la variable aléatoire $D = \dfrac{\sqrt{n}\ (\overline{X} - \mu_0)}{\sigma}$ suit la loi $\mathcal{N}(0 \ ; \ 1)$.

Nous choisirons cette variable aléatoire comme variable discriminante. Remarquons que sa valeur expérimentale $d_e = \dfrac{\sqrt{n}\ (\overline{x} - \mu_0)}{\sigma}$ reflète bien la distance entre $\overline{x}$, la valeur expérimentale du paramètre μ, et μ_0, la valeur à tester.

Définissons $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi } \mathcal{N}(0 \ , \ 1).$$

et la zone de rejet :

$$R_\alpha = \{ \ d \in \mathbb{R} \ / \ |d| > t_{\alpha/2} \ \}$$

La construction du test est achevée.

La mise en œuvre de ce test au niveau de risque α consiste à décider de :

- rejeter l'hypothèse (H) si $\left| \dfrac{\sqrt{n}\ (\overline{x} - \mu_0)}{\sigma} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

<u>Exercice 8-1</u> : On suppose que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B est bruitée et suit une loi normale $\mathcal{N}(\mu \ , \ 2)$. Une personne au point B s'attend à ce que le signal émis ait la valeur 8. Or, le même signal est émis 5 fois du point A, et la valeur moyenne reçue au point B est 9,5. Cette personne doit-elle remettre en cause son hypothèse ?

L'hypothèse $(\mu = \mu_0)$ dont nous venons de décrire le test est ce qu'on appelle une *hypothèse simple*, car, sous cette hypothèse, la loi de l'échantillon est complètement déterminée.
Soit maintenant à tester l'hypothèse *composite* :

$$(H) : \mu \leq \mu_0$$

où μ_0 est une valeur explicite du paramètre.
Construisons-en un test de niveau α ($0 < \alpha < 1$).

On utilise l'estimateur $\bar{X}$ du paramètre μ.

On décidera de rejeter (H) lorsque la valeur de $\bar{x}$ est trop grande par rapport à μ_0, ou, ce qui revient au même, lorsque $d_e = \dfrac{\sqrt{n}\ (\bar{x} - \mu_0)}{\sigma} > c$ pour un certain c.

Pour construire c en fonction du niveau α, supposons que l'hypothèse (H) est vraie, et plus précisément, supposons que μ ($\mu \leq \mu_0$) est la vraie valeur du paramètre. Le risque de rejeter à tort (H) est alors quantifié par $P\{ \dfrac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma} > c\}$. Or, $\dfrac{\sqrt{n}\ (\bar{X} - \mu\ )}{\sigma}$ suit la loi $\mathcal{N}(0, 1)$ et ce risque est donc :

$$P\{ \frac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma} > c\} = P\{ Y > c + \frac{\sqrt{n}\ (\mu_0 - \mu\ )}{\sigma}\ \}$$

où Y suit la loi $\mathcal{N}(0 , 1)$. Il est le plus grand lorsque $\mu = \mu_0$, et il vaut alors $P\{ Y > c \}$. On va donc choisir c tel que cette probabilité soit égale à α. Ainsi, on saura que si l'hypothèse (H) est vérifiée, le test rejettera a tort (H) au plus (environ) $100\ \alpha$ fois sur 100.

En résumé, la mise en œuvre de ce test au niveau de risque α consiste à décider de :

- rejeter l'hypothèse (H) si $\dfrac{\sqrt{n}\ (\bar{x} - \mu_0)}{\sigma} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Y > t_\alpha) = \alpha \quad , \quad Y \text{ suivant la loi } \mathcal{N}(0 , 1).$$

2- Valeur de l'espérance d'une variable normale de variance inconnue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où μ et σ sont inconnus. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) : \mu = \mu_0$$

Construisons-en un test au niveau de risque α ($0 < \alpha < 1$).

On utilise les estimateurs $\overline{X}$ et S de paramètre μ et σ. On note $\overline{x}$ et s leurs valeurs expérimentales.

Si l'hypothèse (H) est vraie, la variable aléatoire $D = \frac{\sqrt{n}\ (\overline{X} - \mu_0)}{S}$ suit la loi de Student à n-1 degrés de liberté. Remarquons qu'il est moins clair que dans le cas où σ est connu que sa valeur expérimentale $\frac{\sqrt{n}\ (\overline{x} - \mu_0)}{s}$ reflète la distance entre $\overline{x}$ et μ_0, puisque le dénominateur s dépend de la valeur expérimentale $(x_1, \ldots, x_n)$. Il est pourtant d'usage de choisir cette variable aléatoire D comme variable discriminante.
Nous définirons alors $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi de Student } t_{n-1} .$$

et la zone de rejet :

$$R_\alpha = \{ d \in \mathbb{R} \ / \ |d| > t_{\alpha/2} \ \}$$

La mise en œuvre du test au niveau de risque α consiste donc à décider de :

- rejeter l'hypothèse (H) si $\left| \frac{\sqrt{n}\ (\overline{x} - \mu_0)}{s} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

On pourrait, de même que dans le cas de la variance connue, construire un test au niveau de risque a de l'hypothèse composite ($\mu \leq \mu_0$).

Exercice 8-2 : L'utilisateur d'un certain câble exige que sa charge moyenne de rupture soit au moins de 200 tonnes. Il a testé 8 de ces câbles et trouvé les charges de rupture :

210 195 197,4 199 198 202 196 195,5

On suppose que la charge de rupture d'un câble suit une loi normale.
Que conclure, au niveau de risque de 5% ? Au niveau de risque de 10% ?

3- Valeur de la variance d'une variable normale

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où μ et σ sont inconnus. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) \ : \ \sigma = \sigma_0$$

Construisons-en un test au niveau de risque α $(0 < \alpha < 1)$.

On utilise les estimateurs $\bar{X}$ et S de paramètre μ et σ. On note $\bar{x}$ et s leurs valeurs expérimentales.

Si l'hypothèse (H) est vraie, la variable aléatoire, $\frac{(n-1)S^2}{\sigma_0^2}$ suit la loi du khi-deux à n-1 degrés de liberté. Nous la choisissons comme variable discriminante. Sa valeur expérimentale $\frac{(n-1)s^2}{\sigma_0^2}$ est fonction de s et permet donc la comparaison de s et σ_0.

On définit la zone de rejet :

$$R_\alpha = \{ d \in \mathbb{R} \ / \ d > t_{\alpha/2} \ \text{ ou } \ d < t_{1-\alpha/2} \ \}$$

avec :

$$P(Y > t_\beta) = \beta \ , \quad Y \text{ suivant la loi } \chi^2_{n-1}$$

La construction du test est achevée.

La mise en œuvre de ce test au niveau de risque α consiste à décider de :

- considérer que l'expérience ne contredit pas (H) si $t_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma_0^2} < t_{\alpha/2}$,
- rejeter l'hypothèse (H) sinon.

En suivant une démarche analogue à celle décrite dans le paragraphe 1, on peut justifier l'utilisation, pour tester l'hypothèse composite ($\sigma \leq \sigma_0$) au risque α, du test qui consiste à

- rejeter l'hypothèse (H) si $\frac{(n-1)s^2}{\sigma_0^2} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon.

Si l'espérance μ est connue, on construit les tests de manière analogue, en utilisant l'estimateur de la variance $S'^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$ et en remarquant que si (H) est vraie, $D = \frac{nS'^2}{\sigma_0^2}$ suit la loi du khi-deux à n degrés de liberté.

Exercice 8-3 : Le système de mesure d'une pompe à essence est tel que le nombre de litres affichés suit une loi normale d'espérance égale au nombre de litres distribués et d'écart-type inconnu σ. Ce système est considéré comme efficace si σ est inférieur à 0,075 litres. Par 20 mesures indépendantes, on a testé un système nouvellement installé et obtenu l'estimation $s^2 = 0{,}00625$. Le système de mesure est-il efficace ?

4- Valeur de la probabilité d'un évènement

Supposons que $X_1, \ldots, X_n$ est un échantillon aléatoire de loi de Bernoulli $\mathfrak{B}(p)$, où p est inconnu.
On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons p_0 , et soit à tester l'hypothèse :

$$(H) \ : \ p = p_0$$

Construisons-en un test au niveau de risque α $(0 < \alpha < 1)$.

On utilise comme estimateur de p la moyenne de l'échantillon, $\bar{X}$. On note p_e la valeur expérimentale correspondante.

Si l'hypothèse (H) est vraie, la variable aléatoire $n\bar{X}$ suit la loi du binôme $\mathfrak{B}(n, p_0)$. Nous la choisissons comme variable discriminante. Par un calcul itératif, - ou en utilisant des tables ou abaques -, on peut déterminer $k^-_{\alpha/2}$ et $k^+_{\alpha/2}$ tels que :

$$k^-_{\alpha/2} = \max \{ k \ / \ \sum_{i=0}^{k} C_n^i \ {p_0}^i (1 - p_0)^{n-i} \leq \alpha/2 \ \}$$

$$k^+_{\alpha/2} = \min \{ k \ / \ \sum_{i=k}^{n} C_n^i \ {p_0}^i (1 - p_0)^{n-i} \leq \alpha/2 \ \}$$

La mise en œuvre de ce test au niveau de risque α consiste alors à décider de :

- considérer que l'expérience ne contredit pas (H) si $k^-_{\alpha/2} < np_e < k^+_{\alpha/2}$,
- rejeter l'hypothèse (H) sinon.

Supposons maintenant que la taille n de l'échantillon est assez grande pour qu'on puisse, sous l'hypothèse (H), donner une bonne approximation de la loi de $\dfrac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}}$ par la loi $\mathfrak{N}(0 ; 1)$. (Il est d'usage de considérer que cette approximation est très bonne lorsque $np_0(1 - p_0) \geq 10$). On peut alors proposer un test au niveau de risque α de mise en œuvre beaucoup plus simple.

On choisit $\dfrac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}}$ comme variable discriminante. Définissant $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi } \mathfrak{N}(0 , 1),$$

on a :

$$P\{ \left| \frac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}} \right| > t_{\alpha/2} \} \approx \alpha \ .$$

La mise en œuvre de ce test au niveau de risque α consiste donc à décider de :

- rejeter l'hypothèse (H) si $\left| \dfrac{\sqrt{n}\ (p_e - p_0)}{\sqrt{p_0(1-p_0)}} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

Exercice 8-4 : La chaîne de fabrication de montres est conçue pour qu'au plus 2% des montres soient défectueuses. Sur 500 montres testées, on en a trouvé 16 défectueuses. Doit-on conclure à un dysfonctionnement de la chaîne de fabrication ? (Proposer et utiliser un test unilatéral).

5- Valeur de l'espérance d'une variable aléatoire de loi quelconque

Supposons que $X_1, \ldots, X_n$ est un échantillon aléatoire de loi quelconque et qu'on veuille tester une hypothèse sur l'espérance μ de sa loi.

Si le type de la loi de l'échantillon est connue, il faut en principe faire une analyse analogue à celle que nous avons faite pour la loi normale : choisir un estimateur ($\bar{X}$ n'est pas forcément le meilleur : voir le chapitre VII §4 ...), choisir une fonction discriminante (de loi connue ou calculable, c'est là le plus gros problème...), etc…

Cependant, pour des valeurs de n assez grandes, et si l'écart-type σ des X_i est connu, on sait, d'après le théorème central-limite, que sous l'hypothèse (H), $\frac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma}$ suit approximativement la loi $\mathcal{N}(0 ; 1)$. On pourra alors construire un test comme on l'a fait dans le paragraphe 4 sur la loi de Bernoulli. Si l'écart-type σ des X_i est inconnu, on utilise généralement la variable discriminante $\frac{\sqrt{n}\ (\bar{X} - \mu_0)}{S}$, en considérant qu'elle suit approximativement la loi $\mathcal{N}(0 ; 1)$, mais on ne peut pas le justifier dans un cadre général. Dans tous les cas, il faut remarquer que si le type de loi des X_i est inconnu, on ne sait pas pour quelles valeurs de n ces approximations sont valides. On ne se risquera pas à utiliser de tels tests si n est plus petit que 30.

6- Intervalle de confiance pour l'estimation d'un paramètre

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi qui dépend d'un paramètre réel inconnu θ, et soit Θ un estimateur du paramètre θ. On suppose disposer d'un échantillon expérimental $(x_1, \ldots, x_n)$, et on note θ_e sa valeur expérimentale. On suppose enfin qu'on dispose d'un test de niveau de risque donné α $(0 < \alpha < 1)$ pour tester les hypothèses $(\theta = \theta_0)$.
On définit alors l'intervalle de confiance au niveau de confiance $(1- \alpha)$ de l'estimation du paramètre θ comme l'ensemble $I_{1-\alpha}$ des valeurs θ_0 qui ne sont pas rejetées par ce test.

En utilisant les tests proposés dans les paragraphes précédents, on obtient les intervalles de confiance au niveau (1- α) suivants :

- Intervalle de confiance de l'espérance d'une variable normale d'écart-type connu σ :

$$I_{1-\alpha} = \left[\bar{x} - t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi } \mathcal{N}(0 , 1).$$

- Intervalle de confiance de l'espérance d'une variable normale de variance inconnue :

$$I_{1-\alpha} = \left[\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi de Student } t_{n-1}.$$

- Intervalle de confiance de la variance d'une variable normale d'espérance inconnue :

$$I_{1-\alpha} = \left[\frac{(n-1)s^2}{t_{\alpha/2}} \quad , \quad \frac{(n-1)s^2}{t_{1-\alpha/2}} \right]$$

avec :

$$P(Y > t_{\beta}) = \beta \qquad \text{où Y suit la loi } \chi^2_{n-1} .$$

- Intervalle de confiance de la variance d'une variable normale d'espérance connue :

$$I_{1-\alpha} = \left[\frac{ns'^2}{t_{\alpha/2}} \quad , \quad \frac{ns'^2}{t_{1-\alpha/2}} \right]$$

avec :

$$P(Y > t_{\beta}) = \beta \qquad \text{où Y suit la loi } \chi^2_{n} .$$

- Intervalle de confiance du paramètre d'une variable de Bernoulli pour les grandes valeurs de n :

$$I_{1-\alpha} \approx \left[p_e - t_{\alpha/2} \sqrt{\frac{p_e(1-p_e)}{n}} \quad , p_e + t_{\alpha/2} \sqrt{\frac{p_e(1-p_e)}{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi } \mathcal{N}(0 , 1).$$

Exercice 8-5 : On suppose que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B est bruitée et suit une loi normale $\mathcal{N}(\mu , 2)$.

a) Pour réduire l'erreur de transmission, on envoie le même signal 9 fois. Les valeurs reçues sont :

5 8,5 12 15 7 9 7,5 6,5 10,5 .

Quel est l'intervalle de confiance bilatéral de la valeur émise μ, au niveau de confiance 0,95 ?

b) Combien de fois le même signal doit-il être envoyé pour que l'intervalle de confiance de μ au niveau 0,95 soit de demi-longueur inférieure à 0,1 ?

Si on dispose d'un test unilatéral de niveau de risque donné α ($0 < \alpha < 1$) pour tester les hypothèses ($\theta \leq \theta_0$), on définit l'intervalle de confiance $]\ldots, +\infty[$ au niveau de confiance $(1- \alpha)$ de l'estimation du paramètre θ comme l'ensemble $I_{1-\alpha}$ des valeurs θ_0 qui ne sont pas rejetées par ce test.

7- Exercices

Exercice 8-6 : Un procédé de fabrication exige d'une certaine solution chimique d'avoir un pH exactement égal à 8,20. La méthode de mesure de pH utilisée donne un résultat qui suit la loi normale d'écart-type 0,02 et d'espérance égale à la vraie valeur du pH. On a mesuré 10 fois le pH de la solution et trouvé :

8,18 8,16 8,17 8,22 8,19 8,17 8,15 8,21 8,16 8,18

a) Que conclure au niveau de risque de 5% ?

b) Que conclure au niveau de risque de 5‰ ?

Exercice 8-7 : On a constaté que sur n = 100 naissances, g = 49 ont été des naissances de garçons. Est-il raisonnable d'admettre que les naissances sont égalements réparties entre garçons et filles ? Même question pour 490 naissances de garçons sur un total de 1 000, de 4 900 sur un total de 10 000.

Exercice 8-8 : Reprendre les données et les questions de l'exercice 8-5, mais en supposant que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B suit une loi normale $\mathcal{N}(\mu , \sigma)$, avec μ et σ inconnus.

Exercice 8-9 : Un procédé de vérification de l'épaisseur de rondelles métalliques fournit une mesure qui suit une loi normale d'espérance égale à la vraie valeur de l'épaisseur et d'écart-type inconnu. On a mesuré 10 fois l'épaisseur d'une rondelle et trouvé :

1,23 1,24 1,26 1,20 1,30 1,33 1,25 1,28 1,24 1,26 mm.

Quel est l'intervalle de confiance au niveau 0,8 de l'écart-type de l'épaisseur d'une rondelle ?

Exercice 8-10 : Dans une population africaine isolée, on a testé 72 personnes choisies au hasard, et observé que 9 d'entre elles portent une anomalie génétique particulière. Quelle est l'intervalle de confiance au niveau 0,95 de la fréquence de cette anomalie dans la population ?

Exercice 8-11 : Entre le premier et second tour des élections présidentielles, un candidat C commande à un institut de sondage une évaluation de ses chances de gagner. Sur 1000 personnes interrogées et ayant l'intention d'exprimer leur suffrage, 515 déclarent avoir l'intention de voter pour C.

a) Si les élections avaient lieu le jour du sondage, C gagnerait-il les élections ? (Proposer un test au niveau de risque de 5%).

b) Le candidat est déçu. Il espérait plus de précision de ce sondage. Combien de personnes ayant l'intention d'exprimer leur suffrage aurait-il fallu interroger pour conclure, au niveau de risque de 5%, que C gagnerait les élections si les élections avaient lieu le jour du sondage ?

Exercice 8-12 : Le diamètre de la prune d'une certaine variété est une variable aléatoire X qu'on suppose normale. Les mesures faites sur un échantillon de 375 prunes de cette variété ont donné les résultats suivants :

diamètre en cm	24	26	28	30	32	34	36	38	40
effectif	7	20	38	79	84	75	53	15	4

a) Estimer, en cm, l'espérance et l'écart-type de X.

b) Quel est l'intervalle de confiance au niveau 0,95 de l'estimation de E(X) ?

IX- Tests portant sur l'égalité des espérances de plusieurs variables aléatoires

1- Egalité des espérances de deux variables normales

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi $\mathcal{N}(\mu_1 ; \sigma_1)$, le deuxième de loi $\mathcal{N}(\mu_2 ; \sigma_2)$, où μ_1 et μ_2 sont inconnus.
On suppose qu'on dispose de valeurs expérimentales $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$ des échantillons, qu'on souhaite utiliser pour tester l'hypothèse :

$$(H) : \mu_1 = \mu_2$$

a) variables normales de variances connues

Supposons σ_1 et σ_2 connus.
Construisons le test.

On utilise les estimateurs $\overline{X}$ et $\overline{Y}$ de μ_1 et μ_2, et on note $\overline{x}$ et $\overline{y}$ les estimations expérimentales correspondantes.

On sait que la loi de $(\overline{X} - \overline{Y})$ est normale, d'espérance $(\mu_1 - \mu_2)$ et d'écart-type $\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$. Sous l'hypothèse (H), la variable aléatoire $D = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ suit donc la loi $\mathcal{N}(0 ; 1)$. Nous la choisissons comme variable discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\frac{|\overline{x} - \overline{y}|}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$$P(Z < -t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha/2 \ , \quad Z \text{ suivant la loi } \mathcal{N}(0 , 1).$$

Exercice 9-1 : Pour mesurer de pH d'une solution, on utilise un pH-mètre qui affiche un résultat dont la loi est $\mathcal{N}(\mu ; 0{,}05)$, où μ est la vraie valeur du pH de la solution. On a mesuré le pH d'une solution A par 12 mesures indépendantes et trouvé une moyenne de 7,04 , et le pH d'une solution B par 10 mesures indépendantes et trouvé une moyenne de 7,05. Peut-on considérer que les deux solutions ont même pH ?

b) variables normales de même variance inconnue

Soit S_1 l'estimateur usuel de σ_1 associé à l'échantillon $X_1, \ldots, X_n$, et notons s_1 l'estimation expérimentale correspondante. Définissons de même S_2 et s_2.

Supposons maintenant les écart-types σ_1 et σ_2 inconnus, mais égaux. Notons σ leur valeur commune.
On peut alors montrer que, sous l'hypothèse (H) ,

$$\frac{(\bar{X} - \bar{Y})}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_1^2 + (m-1)S_2^2}{n + m - 2}}}$$

suit la loi de Student à (n+m-2) degrés de liberté. Nous choisissons cette variable aléatoire comme variable discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\dfrac{|\bar{x} - \bar{y}|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)s_1^2 + (m-1)s_2^2}{n + m - 2}}} > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2$, Z suivant la loi de Student t_{n+m-2} .

Exercice 9-2 : Pour mesurer de pH d'une solution, on utilise un nouveau pH-mètre qui affiche un résultat dont la loi est $\mathcal{N}(\mu ; \sigma)$, où μ est la vraie valeur du pH de la solution et où σ n'a pas été déterminé. On a mesuré le pH d'une solution A par 12 mesures indépendantes et trouvé une moyenne de 7,04 et un écart-type empirique de 0,04 , et le pH d'une solution B par 10 mesures indépendantes et trouvé une moyenne de 7,05 et un écart-type empirique de 0,08. Peut-on considérer que les deux solutions ont même pH ?

c) variables normales de variances inconnues

Si les écart-types σ_1 et σ_2 sont inconnus, et si on n'a pas de raison de les présupposer égaux, on ne peut pas travailler comme dans le paragraphe précédent. En effet, la loi de la fonction discriminante qu'on a proposée dépend alors de la valeur des paramètres inconnus σ_1 et σ_2 et ne peut donc plus être utilisée.
Cependant, si les tailles n et m des échantillons sont très grandes, on pourra considérer que les estimations expérimentales s_1 et s_2 des écart-types sont pratiquement égales à leurs vraies valeurs σ_1 et σ_2, et se ramener ainsi au cas du paragraphe a).
Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consistera alors à :

- rejeter l'hypothèse (H) si $\dfrac{|\bar{x} - \bar{y}|}{\sqrt{\dfrac{s_1^2}{n} + \dfrac{s_2^2}{m}}} > t_{\alpha/2}$,

- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2$, Z suivant la loi $\mathcal{N}(0 , 1)$.

2- Egalité de deux probabilités

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi de Bernoulli $\mathcal{B}(p_1)$, le deuxième de loi $\mathcal{B}(p_2)$, où p_1 et p_2 sont inconnus.
On suppose qu'on dispose des échantillons expérimentaux $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_m)$.

Soit à tester l'hypothèse :

$$(H) : p_1 = p_2$$

Supposons que les tailles n et m des échantillons sont grandes.

Pour construire le test, on utilise les estimateurs classiques $\bar{X}$ et $\bar{Y}$ de p_1 et p_2. On note p_1^e et p_2^e leurs valeurs expérimentales.
Supposons (H) vraie et notons p la valeur commune à p_1 et p_2. Alors, d'après le théorème central-limite, $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p(1-p)}}$ suit une loi proche de $\mathcal{N}(0 ; 1)$.

Cette variable aléatoire ne peut pas être choisie comme fonction discriminante, car le paramètre p est inconnu. Cependant, comme (H) est vraie, $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ est un échantillon aléatoire de taille n+m de loi $\mathcal{B}(p)$, et on peut l'utiliser pour estimer p. Notons p^e la valeur expérimentale de l'estimateur $\dfrac{X_1 + \ldots + X_n + Y_1 + \ldots + Y_m}{n + m}$ de p.

On considère alors que $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p^e(1-p^e)}}$ suit une loi proche de $\mathcal{N}(0 ; 1)$, et c'est cette variable aléatoire qu'on prend fonction discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste alors à :

- rejeter l'hypothèse (H) si $\dfrac{|p_1^e - p_2^e|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p^e(1-p^e)}} > t_{\alpha/2}$,

- considérer que l'expérience ne contredit pas (H) sinon .

où $t_{\alpha/2}$ est défini par :

$$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2 \ , \quad Z \text{ suivant la loi } \mathcal{N}(0 , 1).$$

Ce test ne peut être justifié que si les tailles n et m des échantillons sont grandes. On peut dans le cas contraire utiliser un autre test, celui de Fisher-Irwin, qui est basé sur l'expression des probabilités conditionnelles :

$$P \{ X_1 + \ldots + X_n = i \mid X_1 + \ldots + X_n + Y_1 + \ldots + Y_m = k \}.$$

Exercice 9-3 : Pour mesurer le taux d'occupation d'un matériel, on tire au hasard un echantillon d'instants, et en chacun de ces instants, on regarde si le matériel est ou non occupé. On a obtenu les observations suivantes :

	janvier	février
occupation	400	300
inoccupation	100	100
total	500	400

Les taux d'occupation des mois de janvier et février sont-ils significativement différents ?

3- Egalité des espérances de plusieurs variables normales : méthode de la variance

Soient $X_{i1}, \ldots, X_{in_i}$ (i = 1 à m) m échantillons aléatoires indépendants, de lois normales $\mathcal{N}(\mu_i , \sigma)$ d'espérances $\mu_1, \ldots, \mu_m$ inconnues, et de variance inconnue mais commune σ. On notera $n = \sum_{i=1}^{m} n_i$ le nombre total de variables aléatoires.

On suppose qu'on dispose de valeurs expérimentales $x_{i1}, \ldots, x_{in_i}$ (i = 1 à m) de ces échantillons, qu'on souhaite utiliser pour tester l'hypothèse :

$$(H) : \ \mu_1 = \mu_2 = \ldots = \mu_m$$

Pour construire le test de (H), nous allons proposer deux estimateurs de la variance σ^2, le premier convergeant vers σ^2 que l'hypothèse (H) soit ou non vérifiée, le deuxième ne convergeant vers σ^2 que si (H) est vraie, et, dans le cas contraire, surestimant la valeur de σ^2.

Notons $\bar{X}_i$ et S_i^2 les estimateurs usuels de μ_2 et σ^2 associés à l'échantillon $X_{i1}, \ldots, X_{in_i}$:

$$\bar{X}_i = \frac{X_{i1} + \ldots + X_{in_i}}{n_i} \qquad S_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

On pose :

$$S^2_{intra} = \frac{\sum_{i=1}^{m} (n_i - 1)\, S_i^2}{n - m} \qquad \text{(variance intra-classes)}$$

On sait que $(n_i - 1)\frac{S_i^2}{\sigma^2}$ suit la loi $\chi^2_{n_i}$ et que ces m variables aléatoires sont indépendantes.

Par conséquent, $\frac{(n-m)\, S^2_{intra}}{\sigma^2}$ suit la loi du khi-2 à $\sum_{i=1}^{m} (n_i - 1)$ degrés de liberté, c'est-à-dire la loi χ^2_{n-m}. On a donc :

$$E(\ S^2_{intra}\ ) = \sigma^2$$

S^2_{intra} est donc un estimateur sans biais de σ^2, que l'hypothèse (H) soit ou non vérifiée.

Posons maintenant :

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{m} n_i\ \bar{X}_i}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} n_i} \qquad \text{(moyenne globale)}$$

$$S^2_{inter} = \frac{\sum_{i=1}^{m} n_i\ (\ \bar{X}_i - \bar{\bar{X}}\ )^2}{m - 1} \qquad \text{(variance inter-classe)}$$

On peut montrer par un calcul que :

$$E(\ S^2_{inter}\ ) = \sigma^2 + \frac{1}{m-1} \sum_{i=1}^{m} n_i\ (\ \mu_i - \bar{\mu}\ )^2 \qquad \text{où} \qquad \bar{\mu} = \frac{\sum_{i=1}^{m} n_i \mu_i}{\sum_{i=1}^{m} n_i}$$

Ainsi, si l'hypothèse (H) est fausse, S^2_{inter} n'est pas un estimateur de σ^2, il en surestime sa valeur.

Supposons maintenant (H) vraie. Alors, S^2_{inter} est un estimateur sans biais de σ^2. On peut aussi montrer, mais la preuve n'en est pas élémentaire, que $\frac{(m-1)\, S^2_{inter}}{\sigma^2}$ suit la loi χ^2_{m-1} et que les variables aléatoires S^2_{inter} et S^2_{intra} sont indépendantes. On en conclut que la variable aléatoire $\frac{S^2_{inter}}{S^2_{intra}}$ suit la loi $F_{m-1,\ n-m}$. C'est cette variable qu'on choisit comme variable discriminante. Notons $\frac{s^2_{inter}}{s^2_{intra}}$ sa valeur expérimentale.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste à :

- rejeter l'hypothèse (H) si $\frac{s^2_{inter}}{s^2_{intra}} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Z > t_\alpha) = \alpha \ , \quad Z \text{ suivant la loi } F_{m-1,\ n-m}.$$

Remarque : Dans la pratique, il est souhaitable que les tailles n_i des échantillons soient égales ou presque. Dans ce cas en effet, d'une part on risque moins de considérer comme acceptable l'hypothèse (H) alors qu'elle est fausse, d'autre part, le test est encore relativement bon si les variances des m échantillons ne sont pas tout à fait égales.

Exercice 9-4 : Pour comparer trois types d'essence, on a mesuré la consommation d'essence à vitesse stabilisée de 90km/h de 18 voitures à peu près identiques et obtenu le tableau suivant, où les données sont exprimées en nombre de litres pour 100 km :

essence 1	5,50	6,3	5,95	6,15	6,5	5,6
essence 2	6,1	5,9	6,45	6,05	5,52	5,75
essence 3	6,35	6,8	5,8	5,95	6,4	6,25

La consommation de ces voitures dépend-elle du type d'essence utilisé ?

Exercice 9-5 : Reprendre les données de l'exercice 9-2 avec méthode de la variance. Comparer les conclusions obtenues avec les deux méthodes.

4- Exercices

Exercice 9-6 : On souhaite étudier les effets secondaires d'un certain médicament sur le rythme cardiaque. Pour cela, on a pris le pouls de 11 personnes avant et après la prise de ce médicament, et obtenu les résultats suivants, exprimés en nombre de pulsations par minute :

patient	1	2	3	4	5	6	7	8	9	10	11
avant	74	86	62	98	102	78	64	84	68	79	70
après	70	85	63	90	110	71	60	80	67	69	74

Proposer un test adapté à ces données, en précisant ce qu'il faut supposer pour le justifier. Le mettre en œuvre.

Exercice 9-7 : a) On dispose des notes obtenues à un devoir surveillé par les 24 et 25 étudiants de deux groupes de TD. Quel test proposer pour comparer le niveau de réussite des deux groupes ? Que doit-on supposer pour le justifier ?

b) 10 copies d'examen ont été corrigées par deux correcteurs A et B. Pour chaque copie, on connaît la note donnée par A et la note donnée par B. Quel test proposer pour comparer la sévérité des correcteurs ? Que doit-on supposer pour le justifier ?

Exercice 9-8 : Un constructeur A affirme que la charge de rupture de ses câbles est plus grande que celle des câbles du constructeur B. Pour s'en assurer, un client a fait mesuré la charge de rupture de 14 câbles et trouvé :

câbles A	140	138	143	142	144	137	141	139
câbles B	135	140	136	142	138	140		

Tester l'affirmation du constructeur au risque 0,05.

Exercice 9-9 : Un laboratoire pharmaceutique peut fabriquer un même médicament suivant deux procédés différents, équivalents du point de vue de leur coût. On a mesuré la durée de conservation du médicament par 20 expériences indépendantes et obtenu les durées suivantes, exprimées en nombre d'années :

Procédé A	2,5	3	2	1,5	3,5	1	4	4,5	0,5	2,5
Procédé B	2,2	2,3	2,5	2,8	2,7	2,3	2,8	2,5	2	2,9

Pour chacun des procédés, quels sont les moyenne et écart-type empiriques des résultats? A votre avis, l'un des procédés est-il préférable ?

Exercice 9-10 : Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi $\mathcal{N}(\mu_1 ; \sigma_1)$, le deuxième de loi $\mathcal{N}(\mu_2 ; \sigma_2)$, où μ_1, μ_2, σ_1 et σ_2 sont inconnus. Proposer un test de l'hypothèse ($\sigma_1 = \sigma_2$). Le mettre en œuvre avec les données de l'exercice 9-2.

X- Tests d'hypothèses non-paramétriques sur la loi d'une variable aléatoire

Exemple 10-1 : On a lancé un dé 360 fois et obtenu le tableau :

n° de la face	1	2	3	4	5	6
effectif	43	55	51	71	72	68

Comment utiliser ces données pour tester l'hypothèse que toutes les faces ont la même probabilité ? •

Exemple 10-2 : Dans les exemples des deux derniers chapitres, nous avons souvent supposé que la loi d'un échantillon dont on disposait d'une valeur expérimentale suivait une loi normale. Comment tester une telle hypothèse ? Supposons par exemple que Z_1, …, Z_n est un échantillon aléatoire de loi inconnue, et que nous voulons tester l'hypothèse :

$$(H) : \text{ la loi de } Z_1, \ldots, Z_n \text{ est } \mathcal{N}(0 , 1).$$

à l'aide de valeurs expérimentales z_1, …, z_n.

Contrairement au cas précédent, la loi de référence est ici continue. On se ramène au cas discret en découpant l'ensemble des valeurs possibles des variables aléatoires Z_1, …, Z_n en un nombre fini k de régions, en général des intervalles, R_1, …, R_k. Les données expérimentales z_1, …, z_n se répartissent suivant le tableau d'effectifs :

région	R_1	R_2	……	R_k
effectif	c_1	c_2	……	c_k

Si l'hypothèse (H) est vraie, on sait calculer la probabilité p_a pour qu'un résultat Z tombe dans la zone R_a. On est donc ramené à une situation analogue à celle de l'exemple 10-1. Il faudra cependant être plus prudent dans l'interprétation du résultat du test, car il peut dépendre de la manière dont les régions R_a ont été délimitées. •

1- Egalité de la loi de l'échantillon et d'une loi spécifiée

Soit Y_1, …, Y_n un échantillon aléatoire à valeur dans {1, 2, …, k} de loi inconnue. Pour simplifier la présentation, notons Y une variable aléatoire de même loi. Nous supposons disposer d'une valeur expérimentale y_1, …, y_n de l'échantillon, et nous voulons l'utiliser pour tester l'hypothèse :

$$(H) : \forall\ a \in \{1, 2, \ldots, k\} \quad P\{ Y = a \} = p_a$$

où les probabilités p_a sont données et vérifient $\sum_{a=1}^{k} p_a = 1$.

Pour a dans {1, 2, …, k}, posons :

$$C_a = \text{card } \{ i \ / \ Y_i = a \}$$

et notons c_a la valeur expérimentale correspondante.

Sous l'hypothèse (H), C_a est une variable aléatoire de loi $\mathfrak{B}(n , p_a)$. Son espérance est np_a . La valeur prise par $(C_a - np_a)^2$, lorsque n est grand, donne donc une indication de la plausibilité de l'hypothèse que C_a est une variable aléatoire de loi $\mathfrak{B}(n , p_a)$: plus cette valeur est grande, moins cette hypothèse est plausible.
De fait, on choisit comme fonction discriminante :

$$D = \sum_{a=1}^{k} \frac{(C_a - np_a)^2}{np_a}$$

et on décidera de rejeter (H) lorsque la valeur expérimentale d de D est trop grande.

Remarque : Dans le contexte d'utilisation de ce test, la valeur prise par Y n'intervient que comme un outil pour classer les individus de la population étudiée. La fonction discriminante D est définie à partir des contingents des différentes classes de l'échantillon observé. Y pourrait tout autant être une variable aléatoire qualitative, comme dans l'exemple 10-2, au lieu d'être numérique. •

a) Test du khi-deux

On peut montrer, - mais, dès que k est plus grand que 2, la preuve n'est pas élémentaire -, que si (H) est vraie et si n est grand, D suit approximativement la loi χ^2_{k-1}.
Dans la pratique, on utilise cette approximation si pour tout a, $np_a \geq 1$ et si pour au moins 80% des a, $np_a \geq 5$.

La mise en œuvre du test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\sum_{a=1}^{k} \frac{(c_a - np_a)^2}{np_a} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Z > t_\alpha) = \alpha \ , \quad Z \text{ suivant la loi } \chi^2_{k-1}$$

Exercice 10-1 : a) Les données sont celles de l'exemple 10-1. Tester l'hypothèse que toutes les faces ont la même probabilité, au niveau de risque de 2%, puis au niveau de risque de 5%.
b) Supposer toutes les effectifs multipliés par 2, et tester la même hypothèse.

b) Test par simulation

Notons encore d la valeur expérimentale de D.

Si la taille de l'échantillon ne permet pas l'approximation de la loi de la fonction discriminante D par une loi du khi-deux, on peut utiliser une simulation de cette loi sur ordinateur :

- On tire indépendamment n valeurs y suivant la loi de Y donnée par l'hypothèse (H), et on calcule la valeur de D correspondante. Notons-la d_1.
- On recommence un grand nombre r de fois ce tirage. On obtient les valeurs $d_1, \ldots, d_r$.
- De la loi des grands nombres, on déduit que, sous l'hypothèse (H) :

$$P\{ D \geq d \} \approx \frac{\text{card}\{ i \ / \ d_i \geq d \}}{r}$$

La mise en œuvre du test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\frac{\text{card}\{ i \ / \ d_i \geq d \}}{r} < \alpha$,
- considérer que l'expérience ne contredit pas (H) sinon .

2- Cas où certains paramètres ne sont pas spécifiés

Exemple 10-3 : Reprenons l'exemple 10-2, mais supposons maintenant à tester l'hypothèse :

(H) : la loi de $Z_1, \ldots, Z_n$ est normale.

Sous cette seule hypothèse, les probabilités p_a pour qu'un résultat Z tombe dans la zone R_a ne sont pas calculables. Pour tester (H_Z), on estime les paramètres μ et σ de la loi de l'échantillon $Z_1, \ldots, Z_n$ par les estimateurs usuels $\overline{X}$ et S. On teste ensuite l'hypothèse :

(H') : la loi de $Z_1, \ldots, Z_n$ est $\mathcal{N}(\overline{x}, s)$

comme dans le paragraphe précédent, soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k - 1- e), où e est le nombre de paramètres estimés (ici, e=2).•

Exercice 10-2 : On a relevé le nombre d'accidents durant une période de 30 semaines dans un secteur donné, et obtenu :

8	0	0	1	3	4	0	2	12	5
1	8	0	2	0	1	9	3	4	5
3	3	4	7	4	0	1	2	1	2

Peut-on considérer que ce nombre suit une loi de Poisson ? (On utilisera la partition de l'ensemble des valeurs possibles : {0} {1} {2, 3} {4, 5} {6 ou plus}).

3- Egalité des lois de plusieurs échantillons

Soient $Y_{i1}, \ldots, Y_{in_i}$ (i = 1 à m) m échantillons aléatoires indépendants de lois inconnues, toutes les variables aléatoires prenant leurs valeurs dans {1, 2, …, k}, et soit à tester l'hypothèse (H) :

(H) : Les lois des m échantillons sont identiques

Notons $y_{i1}, \ldots, y_{in_i}$ les valeurs expérimentales des échantillons.

Pour simplifier la présentation, notons, pour tout i, Y_i une variable aléatoire ayant la même loi que l'échantillon $Y_{i1}, \ldots, Y_{in_i}$. Avec cette notation, (H) se réécrit :

$$(H) : \quad \forall\ a \in \{1, 2, \ldots, k\} \quad P\{ Y_1 = a \} = \ldots = P\{ Y_m = a \}$$

Supposons d'abord l'hypothèse (H) vraie. Notons alors Y une variable aléatoire ayant la même loi que les Y_i. Pour a dans {1, 2, …, k}, estimons les probabilités P{ Y = a } par :

$$p_a = \frac{\text{card}\{ (i,j) \ / \ y_{ij} = a \}}{\sum_{i=1}^{m} n_i}$$

Posons :

$$(H') \ : \forall\ a \in \{1, 2, \ldots, k\} \quad \forall\ i \in \{1, 2, \ldots, m\} \quad P\{ Y_i = a\} = p_a$$

On teste (H') par une méthode semblable à celle du paragraphe 1. On définit pour cela :

$$D = \sum_{i=1}^{m} \sum_{a=1}^{k} \frac{(C_{ia} - n_i p_a)^2}{n_i p_a}$$

où :

$$C_{ia} = \text{card}\{ j \ / \ Y_{ij} = a \}$$

et on décide de rejeter (H), lorsque la valeur expérimentale d de D est trop grande.

On procède soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k-1)(m-1).

Exercice 10-3 : On a testé trois modèles de machines à laver, A, B et C, en comptant le nombre de pannes durant leur 3 premières années de fonctionnement. On a obtenu le tableau :

	0 panne	1 panne	2 pannes	3 pannes ou plus
A	884	403	95	23
B	123	693	373	28
C	57	219	144	8

Cette expérience met-elle en évidence une différence entre les trois modèles ?

4- Indépendance de deux caractères aléatoires

On étudie conjointement deux caractères des individus d'une population, qui prennent leurs valeurs respectivement dans {1, 2, …, k} et {1, 2, …, m}. On suppose disposer de n valeurs expérimentales indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$.

Pour représenter cette situation, on introduit (X_1,Y_1), …, (X_n,Y_n) un échantillon de n variables aléatoires indépendantes à valeurs dans $\{1, 2, \ldots, k\}\times\{1, 2, \ldots, m\}$ de même loi (inconnue). Notons (X,Y) une variable aléatoire de même loi. Quels que soient les couples (a_i, b_i) on a donc, par hypothèse :

$$P\{ [(X_1,Y_1) = (a_1, b_1)] \text{ et } \ldots \text{ et } [(X_n,Y_n) = (a_n, b_n)] \} =$$
$$= P\{ (X_1,Y_1) = (a_1, b_1) \} \ldots P\{ (X_n,Y_n) = (a_n, b_n) \} =$$
$$= P\{ (X,Y) = (a_1, b_1) \} \ldots P\{ (X,Y) = (a_n, b_n) \}$$

Soit à tester l'hypothèse d'indépendance des caractères, autrement dit l'hypothèse :

$(H) : \forall (a, b)\in \{1, 2, \ldots, k\}\times\{1, 2, \ldots, m\}\ P\{ (X,Y) = (a, b) \} = P\{ X = a \}P\{ Y = b \}$

On estime les lois (marginales) de X et Y par :

$$p_a^X = \frac{card\{ i \ / \ x_i = a \}}{n} \qquad a \in \{1, 2, \ldots, k\}$$

$$p_b^Y = \frac{card\{ i \ / \ y_i = b \}}{n} \qquad b \in \{1, 2, \ldots, m\}$$

On choisit comme fonction discriminante :

$$D = \sum_{a=1}^{k} \sum_{b=1}^{m} \frac{(C_{(a,b)} - np_a^X p_b^Y)^2}{np_a^X p_b^Y}$$

où :

$$C_{(a,b)} = card\{ i \ / \ (X_i, Y_i) = (a, b) \}$$

et on décide de rejeter (H) lorsque la valeur expérimentale d de D est trop grande.

On procède soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k-1)(m-1).

Exercice 10-4 : On a interrogé 2000 personnes lors de leur départ en vacances sur leur destination et le moyen de transport utilisé pour s'y rendre. On a obtenu le tableau :

	Campagne	Mer	Montagne
Voiture	250	700	350
Train	200	200	50
Avion	15	200	35

Y a-t-il un lien entre la destination et le moyen de transport ?

5- Test des signes

Exemple 10-4 : On a testé un médicament contre l'hypertension sur 18 patients en mesurant la différence entre leur tension avant le début du traitement et après un mois de traitement. On a obtenu les résultats :

-2	-1	+1	+3	-8	+1	+2	-4	-5
-3	-3	-6	-2	-7	+2	-7	-5	-4

On se demande si le médicament a un effet réel sur l'hypertension, - ou s'il est efficace contre l'hypertension -.
Notons X la variable aléatoire qui représente cette différence.
- Si on peut supposer que la loi de X est normale, on peut utiliser un test de Student de l'hypothèse simple "l'espérance de X est nulle" ou de l'hypothèse composite "l'espérance de X est négative". (Le fait que la loi de X peut être considéré comme normale peut lui-même être testé par un test du khi-deux, mais le test sera ici grossier car l'effectif total est faible.)
- Si on ne peut pas supposer la loi normale, on peut proposer de tester une l'hypothèse sur la valeur de sa médiane. •

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi inconnue, de médiane m. On note F sa fonction de répartition, qu'on suppose pour simplifier continue.
Soit à tester l'hypothèse :

$$(H) : \quad m = m_0$$

où m_0 est un réel spécifié.
Introduisons les variables aléatoires (indépendantes) Y_i :

$$Y_i = 1 \quad \text{si} \quad X_i \leq m_0 ,$$
$$= 0 \quad \text{sinon} .$$

Elles suivent la loi de Bernoulli $\mathcal{B}(F(m_0))$. L'hypothèse (H) équivaut donc à l'hypothèse :

$$(H') : \text{ le paramètre de la loi de Bernoulli des } Y_i \text{ vaut } \frac{1}{2} .$$

On est ramené au cas traité dans paragraphe 4 du chapitre VIII.
On pose donc :

$$D = \text{card}\{ i \ / \ X_i \leq m_0 \}$$

et on note d sa valeur expérimentale.

Le test consiste à :
- rejeter (H) si $\alpha > 2\, P\{ Z < \min(d, n-d) \}$,
- considérer que l'expérience ne contredit pas (H) sinon,

où Z suit la loi $\mathcal{B}(n , \frac{1}{2})$.

Exercice 10-5 : a) Utiliser le test des signes pour traiter l'exemple 10-4.
b) Remarquer qu'on peut aussi tester l'hypothèse (H') du test des signes par un test du khi-deux. Que trouve-t-on par cette méthode ?
c) Que conclut-on en utilisant un test de Student ?

6- Exercices

Exercice 10-6 : Reprendre les données de l'exercice 8-12, et tester la normalité de la loi de X.

Exercice 10-7 : Proposer une deuxième façon de traiter l'exercice 9-3.

Exercice 10-8 : Sur 100 tubes à vide testés, 41 ont eu une durée de vie de moins de 30 heures, 31 entre 30 et 60 heures, 13 entre 60 et 90 heures, et 15 plus de 90 heures. Ces données sont-elles compatibles avec l'hypothèse que la durée de vie d'un tube à vide est une loi exponentielle d'espérance égale à 50 heures ?

Exercice 10-9 : Le tableau ci-dessous donne la répartition de 200 naissances en fonction de la parité de la mère et du poids du nouveau-né.

	primipares	multipares
poids inférieur à 3kg	26	20
entre 3 et 4 kg	61	63
supérieur à 4 kg	8	22

Les deux caractères, parité de la mère et poids du nouveau-né, sont-ils statistiquement reliés ?

Exercice 10-10 : 2000 personnes ont passé un concours. Proposer une méthode de comparaison des manières de noter de deux correcteurs A et B, sachant qu'on peut pour cela demander à chacun de corriger 50 copies.

Exercice 10-11 : Reprendre l'exercice 9-6 en utilisant le test des signes.

IUP2-MIAGE 2000-2001

Bordeaux I

Devoir surveillé du 17-11-00

Documents et calculettes autorisés.
Le soin apporté à la rédaction sera apprécié.
Toutes les réponses doivent être argumentées.

I

Un stock important comprend 40% de transistors de type A, 60% de type B.
Exprimée en heures d'utilisation, la durée de vie d'un transistor de type A suit la loi exponentielle de paramètre a=1. La durée de vie d'un transistor de type B suit la loi exponentielle de paramètre b=2.
On prend au hasard un transistor dans le stock. On note D sa durée de vie.

1) Que vaut la probabilité $P(D \geq 2)$?

2) a) Quelle est la fonction de répartition de D ? Est-elle continue en tout point de $\mathbb{R}$?
b) La loi de D est-elle à densité ? Si oui, quelle est cette densité ?
c) Calculer E(D).

3) On constate que le transistor qu'on a tiré fonctionne toujours au bout de deux heures d'utilisation. Avec quelle probabilité est-il du type A ?

4) On tire au hasard dans le stock 5 transistors. Avec quelle probabilité 2 d'entre eux exactement sont-ils du type A ?

II

On sait que les pommiers d'une plantation récente sont porteurs d'un certain virus avec la probabilité p, indépendamment les uns des autres. Les pommiers atteints seront contagieux dans un an. On décide de faire analyser les pommiers pour détruire à temps ceux qui sont porteurs du virus.

A- Un laboratoire de virologie est chargé d'analyser la sève de 10 pommiers pris au hasard dans la plantation, et de conclure pour chaque pommier s'il est ou non porteur du

virus. Le laboratoire dispose d'un test très fiable mais coûteux qui permet de détecter la présence du virus dans un échantillon de sève, et quelle qu'en soit sa concentration. Plutôt que d'analyser un par un les 10 échantillons, il utilise la méthode (M) suivante :

Méthode (M) : Après avoir mis de côté et étiqueté la moitié de chacun des 10 échantillons, mélanger les 10 demi-échantillons restants et analyser ce mélange. Si le virus n'y est pas détecté, aucun des 10 pommiers n'est porteur du virus. Sinon, analyser séparément chacun des 10 demi-échantillons qu'on avait mis de côté.

On note N le nombre (aléatoire) d'analyses effectuées.

a) Quelles sont les valeurs que peut prendre N ?

b) Montrer que N peut s'écrire $N = 1 + 10\,X$, où X est une variable de Bernoulli de paramètre α à déterminer.

c) Que valent, en fonction de α, $E(N)$ et $\sigma(N)$?

d) Pour quelles valeurs de p a-t-on $E(N) < 10$?

B- Le laboratoire a en fait à analyser les prélèvements issus de 10n pommiers. Il les répartit en n lots de 10 et utilise pour chaque lot la méthode (M).

On note T le nombre total d'analyses effectuées pour conclure pour chaque pommier s'il est ou non porteur du virus.

1) Ecrire T comme une somme de variables aléatoires.

2) On suppose dans cette question $n=400$ et $p=0{,}01$.

a)Vérifier que $E(N) \approx 1{,}96$ et $\sigma(N) \approx 2{,}94$.

b) Donner une valeur approchée de $P(\{ T > 1000 \})$.

Commenter ce résultat, quant à la comparaison de la méthode (M) et de la méthode banale qui consiste à analyser les 4000 échantillons.

3) On revient au cas général.

a) Soit r un réel strictement plus grand que E(N). Montrer que si n est "très grand" :

$$P\left(\frac{T}{n} < r \right) \approx 1 .$$

b) On remarque que la méthode (M) est préférable à la méthode banale lorsque $T < 10n$. On suppose n "très grand". Pour quelles valeurs de p la méthode (M) est-elle, avec une probabilité proche de 1, préférable à la méthode banale ?

Probabilités et statistiques : examen du 22-01-01

Durée conseillée : 1h20

Documents et calculettes autorisés.

Le soin apporté à la rédaction sera apprécié.

Toutes les réponses doivent être argumentées.

I

Les tabourets de cafétéria fabriqués dans l'usine A sont défectueux avec la probabilité p_A, ceux fabriqués dans l'usine B le sont avec la probabilité p_B ($p_A \neq p_B$). Un client, qui ne connaît pas les valeurs de p_A et p_B, tire à pile ou face l'une des usines, puis y commande 50 tabourets. On note N le nombre de tabourets défectueux qui lui sont livrés.

1) Décrire la loi de N ?

2) Soit p est un réel dans [0, 1] et n un entier positif. Interpréter la somme :

$$\sum_{k=0}^{n} k\ C_n^k\ p^k\ (1-p)^{n-k}$$

et en déduire, sans calcul, sa valeur.

3) Quelle est l'espérance de N ?

II

Une machine fabrique des vis dont la longueur X est une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma)$. On a mesuré la longueur de 100 vis prises au hasard dans la production et obtenu les résultats suivants :

longueur (en mm)	31	32	33	34	35
effectif	6	21	38	25	10

1) Définir un échantillon aléatoire adapté à l'énoncé. Dans ce cadre modèle, comment s'interprète le premier effectif, 6, du tableau ?

2) Calculer les estimations $\bar{x}$ et s de l'espérance et l'écart-type de la loi de X. (On précisera les estimateurs choisis ainsi que les expressions calculées).

3) Avant de faire la série de mesures, la machine avait été réglée pour que l'espérance μ soit égale à 33,5. Doit-on penser que la machine s'est déréglée ? (On précisera la variable aléatoire discriminante choisie, sa loi, etc... et on justifiera ce choix).

4) Les résultats des 100 mesures confirment-ils ou infirment-ils le fait que la loi de l'échantillon est normale ? (Les mêmes précisions qu'en 3) sont demandées. On utilisera le tableau suivant, où Π représente la fonction de répartition de la loi normale $\mathcal{N}(0,1)$:

u	31,5	32,5	33,5	34,5
$\Pi\left(\frac{u - \bar{x}}{s}\right)$	0,061	0,277	0,642	0,906

IUP2-MIAGE 2000-2001

Bordeaux I

Examen du 2-7-01

Durée : 2h

Documents et calculettes autorisés.
Le soin apporté à la rédaction sera apprécié.
Toutes les réponses doivent être argumentées.

I

Le rayon "télévision" d'un magasin d'une petite ville propose 2 modèles A et B. On a constaté qu'un visiteur de ce rayon achète un poste de la marque A avec la probabilité p_A, un poste de la marque B avec la probabilité p_B, n'achète rien avec la probabilité q ($p_A > 0$, $p_B > 0$, $q > 0$, $p_A + p_B + q = 1$), et que les choix des visiteurs sont indépendants.
100 personnes visitent ce rayon.

1) Quelle est la probabilité ρ de l'évènement "les 10 premiers visiteurs achètent un poste A, les 20 suivants un poste B, les 70 autres n'achètent rien" ?

2) On note X le nombre de clients qui achètent un poste A, Y le nombre de clients qui achètent un poste B.
a) Quelle est la loi de X ?
b) Que vaut E(X) ?
c) Exprimer $P\{ X = 60 \}$.

3) a) Préciser quel est l'évènement { X = 60 et Y = 60}. En déduire la valeur de P{ X = 60 et Y = 60}. Les variables aléatoires X et Y sont-elles indépendantes ?
b) A-t-on : $P\{ X = 10 \text{ et } Y = 20\} = C_{100}^{10}\ C_{90}^{20}\ {p_A}^{10}\ {p_B}^{20}\ q^{70}$? (Justifier la réponse).

4) Le vendeur fait un bénéfice de α francs sur la vente d'un poste A, de β francs sur la vente d'un poste B. On note T le bénéfice correspondant aux 100 visiteurs.
a) Exprimer T en fonction de X et Y.
b) Que vaut E(T) ?
c) Peut-on affirmer : $\nu(T) = \alpha^2\, \nu(X) + \beta^2\, \nu(Y)$? (Justifier la réponse).

II

Dans un laboratoire, on a effectué 100 expériences indépendantes pour mesurer une certaine grandeur g. 60 expériences ont été menées dans les conditions A, 40 l'ont été dans les conditions B. On a obtenu les résultats suivants :

conditions A :

g	16	18	19	20	21	22	23	24
effectif	1	6	16	19	8	7	2	1

conditions B :

g	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	28	31
effectif	2	2	1	2	2	3	2	6	2	4	3	1	2	3	3	1	1

1) Que conclure de ces résultats, quant à l'influence des conditions d'expérience, A ou B, sur la mesure de g. (Décrire et mettre en œuvre un test du khi-deux. Pour éviter des effectifs de classes trop petits, on groupera les valeurs de g suivant la partition : $\{ g \leq 18 \}$, $\{ 19 \leq g \leq 20 \}$, $\{ g \geq 21 \}$).

2) On suppose que dans les conditions A, la mesure de g est une variable aléatoire X d'espérance μ_A et d'écart-type σ_A, et que dans les conditions B, la mesure de g est une variable aléatoire Y d'espérance μ_B et d'écart-type σ_B.

a) Donner les estimations $\bar{x}$ et s_A de μ_A et σ_A. (On précisera les estimateurs choisis ainsi que les expressions calculées).

On trouverait de même les estimations $\bar{y}$ et s_B de μ_B et σ_B :

$$\bar{y} = 20 \qquad s_B \approx 4{,}56$$

b) Les résultats expérimentaux confirment-ils ou infirment-ils l'hypothèse (H_μ) :

$$(H_\mu) : \quad " \mu_A = \mu_B "$$

c) Les résultats des tests du 1) et du 2-b) sont-ils contradictoires ? Peut-on imaginer répondre à la question 1) par un autre test que celui du khi-deux ? (... on ne demande ici ni de construire ni de mettre en œuvre un tel test...).

IUP2-MIAGE 2001-2002

Bordeaux I

Examen du 25-01-02

Durée : 2h

Documents et calculettes autorisés.
Le soin apporté à la rédaction sera apprécié.
Toutes les réponses doivent être argumentées.

I

Un joueur a dans sa poche 3 pièces d'apparences identiques :
- la pièce A est normale et parfaitement équilibrée,
- la pièce B est truquée : si on la lance, on obtient "face" avec la probabilité 2/3, "pile" avec la probabilité 1/3,
- la pièce C est aussi truquée : si on la lance, on obtient "face" avec la probabilité 1/3, "pile" avec la probabilité 2/3.

1) Le joueur a pris au hasard une pièce dans sa poche. Il l'a lancée une première fois et a obtenu "face". Il l'a relancée et a obtenu "face". Il l'a lancée une troisième fois et a obtenu "pile". Avec quelle probabilité est-ce la pièce A qu'il a lancée ?

2) 36 fois de suite, le joueur prend une pièce au hasard, la lance, et la remet dans sa poche. On note N le nombre de faces qu'il obtient.

a) Quelle est la loi de N ?

b) Quelle est l'espérance de N ?

II

1) Exprimée en heures d'utilisation, la durée de vie X d'une aiguille de machine à coudre suit une loi uniforme sur l'intervalle [10, 30].

a) On note α le nombre tel que, avec une probabilité égale à 0,9 , X est plus grande que α . Calculer α .

b) Que valent l'espérance et l'écart-type de X ?

2) L'utilisateur d'une machine à coudre a acheté 25 aiguilles, dont on suppose les durées de vie indépendantes et de loi uniforme sur l'intervalle [10, 30]. On note Y la durée totale d'utilisation de la machine à coudre que ces aiguilles permettent. (... une et une seule aiguille est nécessaire au fonctionnement de la machine...).

a) Calculer l'espérance et l'écart-type de Y.

b) On note β le nombre tel que, avec une probabilité égale à 0,9 , Y est plus grande que β . Donner une valeur approchée de $\beta / 25$.

III

On a mesuré les durées de vie de 100 objets produits dans les mêmes conditions et obtenu :

durée de vie	1	3	5	7	9	11	13	15
nombre d'objets	17	18	19	11	11	11	8	5

a) Définir un échantillon aléatoire adapté à l'énoncé.

b) Donner une estimation, qu'on notera m, de l'espérance de cet échantillon.(On précisera l'estimateur choisi ainsi que l'expression calculée).

c) Peut-on supposer que la loi de l'échantillon est une loi exponentielle ?

(On utilisera le tableau :

a	2	4	6	8	10	14
$e^{-a/m}$	0,732	0,536	0,393	0,288	0,217	0,113

On rappelle que quels que soient les réels a et b, $\int_a^b \frac{1}{m} e^{-x/m} dx = e^{-a/m} - e^{-b/m}$)

Réponses aux exercices des fins de chapitres

1-5 : $P(A)+ P(B) - P(A \cap B) = P(A \cup B) \leq 1$

1-6 : $\frac{C_{20}^2}{C_{30}^2} \approx 0{,}437$ $\frac{C_{10}^2}{C_{30}^2} \approx 0{,}103$

1-7 : a) $\frac{1}{C_n^2} = \frac{2}{n(n-1)}$ b) $\frac{n-r}{C_n^2} = \frac{2(n-r)}{n(n-1)}$

1-8 : a) $\frac{C_{10}^2}{C_{20}^4} \approx 0{,}0093$ b) $1 - \frac{2^4 C_{10}^4}{C_{20}^4} \approx 0{,}31$

2-7 : $1 - (\frac{1}{2})^3 = 0{,}875$

2-8 : 0,59 (0,03 pour 50 personnes)

2-9 : 1) $P = (\frac{2}{3})^{k-1}\frac{1}{3}$ si $k \geq 1$) 2) $P = \frac{1}{3}$ si $k = 1$, $P = (\frac{1}{2})^{k-2}\frac{1}{3}$ si $k \geq 2$

3) $P = \frac{1}{3}$ si $k = 1$, 2 ou 3 .

2-10 : a) ≈ 0,84 b) ≈ 0,9995 (a) ≈ 0,33 b) ≈ 0,99995 si 5‰ malades)

2-11 : a) $p_1p_2...p_n$ b) $1- (1-p_1)(1-p_2)...(1-p_n)$ c) $1- (1-p_1p_2)(1-p_3p_4)$

d) $p_5 (1- (1-p_1)(1-p_3)) (1- (1-p_2)(1-p_4)) + (1-p_5) (1- (1-p_1p_2)(1-p_3p_4))$

2-12 : a) $p^2 + (1-p)^2$ b) $(p^2 + (1-p)^2)^n$ c) $\frac{1}{(1 + (\frac{p}{1-p})^2)^n}$

2-13 : Proba de gagner = a) 1/3 b) 2/3 c) 1/2 . Stratégie b préférable.

3-10 : $C_5^2 p^2 (1-p)^3$ avec $p = \int_{200}^{300} \frac{200}{x^2}\,dx = 1/3$

3-11 : $P(N_1 = 1 \text{ et } N_2 = 1) = 3/10$ $P(N_1 = 1 \text{ et } N_2 = 2) = P(N_1 = 2 \text{ et } N_2 = 1) = 1/5$

$P(N_1 = 1 \text{ et } N_2 = 3) = P(N_1 = 2 \text{ et } N_2 = 2) = P(N_1 = 3 \text{ et } N_2 = 1) = 1/10$

3-12 : a) $F(x) = 0$ si $x < 0$, $F(x) = x^n$ si $0 \leq x < 1$, $F(x) = 1$ si $x \geq 1$

$f(x) = 0$ si $x < 0$ ou $x \geq 1$, $f(x) = nx^{n-1}$ si $0 \leq x < 1$

b) $f(x) = 0$ si $x < 0$ ou $x \geq 1$, $f(x) = n(1-x)^{n-1}$ si $0 \leq x < 1$

4-11 : $E(M) = \frac{n}{n+1}$ $\nu(M) = \frac{n}{(n+2)(n+1)^2}$

4-12 : $P(N=n) = p^{n-1} (1-p)$ ($n \geq 1$). $E(N) = \frac{1}{1-p}$.

4-13 : 1) a) $(c_u+t_u) \frac{n}{1-p}$ b) $(nc_u+t_m) \frac{1}{(1-p)^n}$

2) a préférable à b si $p \geq 1 - (\frac{2n+1}{3n})^{1/(n-1)} \approx 0{,}039$ si n=10, ≈ 0,004 si n=100.

5-5 : $a = 1 - (0,99)^{10} - 10\,(0,99)^9\,0,01 \approx 0,0043$; $3a(1-a)^2 \approx 0,01269$;
$1-(1-a)^3 \approx 0,01274$.

5-6 : $p = 0,1$ $\lambda = 4$. $P(C=k) = \sum_{n\geq k} C_n^k\, p^k\,(1-p)^{n-k}\, e^{-4}\,\frac{4^n}{n!} = e^{-4p}\,\frac{(4p)^k}{k!}$. C suit $\mathcal{P}(4p)$.

5-7 : a) $\sigma = 1/1,645 \approx 0,608$ b) $s = 1,96\,\sigma \approx 1,19$

5-8 : 0,383 .

6-2 : $n \geq 790$ (à comparer à $n \geq 2917$)

6-3 : au moins 26549 F .

6-4 : $s \geq 1,96\sqrt{p(1-p)}\sqrt{n}$. Si $n = 100$, $s \geq 10$, $2s/(n+2s) \approx 17/100$. Si $n = 1000$, $s \geq 31$, $2s/(n+2s) \approx 6/100$.

6-5 : a) $\mathcal{B}(v, p)$. $E(Y) = vp$. $v(Y) = vp(1-p)$

b) $R = xY$. $E(R) = xvp = xve^{-cx}$. E(R) maxi si $x = 1/c$ et alors, $E(R) = v\,\frac{e^{-1}}{c}$.

c) $n \geq 0,85\sqrt{vp(1-p)} + vp$.

8-6 : $\alpha_{crit} = 0,0096$ a) $pH \neq 8,2$ b) " $pH = 8,2$ " n'est pas contredite par l'expérience.

8-7 : (H) " Proba (garçon) = 0,5 "

a) $\alpha_{crit} = 0,84$ (H) acceptable b) $\alpha_{crit} = 0,53$ (H) acceptable c) $\alpha_{crit} = 0,046$ (H) rejetée au risque de 5% , (H) acceptable au risque de 4%.

8-8 : a) $I_{0,95} \approx [\ 6,63\ ,\ 11,37\ ]$ b) 1537

8-9 : $I_{0,9} \approx [\ 0,029\ ,\ 0,054\ ]$

8-10 : $I_{0,95} \approx [\ 0,048\ ,\ 0,20\ ]$

8-11 : a) " $p < 0,5$ " compatible avec l'expérience b) $\frac{\sqrt{n}\ 0,015}{\sqrt{1/4}} \approx 1,65$ $n \geq \approx 3000$

8-12 : a) $\bar{x} = 31,99$; $s = 3,29$ b) $I_{0,95} = [\ 31,66\ ,\ 32,33\ ]$

9-6 : X = pouls avant, Y = pouls après. X et Y non indépendantes. On suppose loi de Z = Y-X normale. " $E(Z) = 0$ " n'est pas contredite par l'expérience. (test de Student au risque de 5% (et jusqu'à $\alpha_{crit} \approx 0,18$)). Effet secondaire non démontré.

9-7 : a) Test ch9§1c. Supposer normalité des notes du groupe A, du groupe B, indépendance des échantillons. b) cf. exo 9-6.

9-8 : Supposer normalité des échantillons et même variance. " $\mu_A < \mu_B$ " non rejetée par test unilatéral t_{12} au risque de 5% ($\alpha_{crit} \approx 0,09$) . L'affirmation du constructeur A n'est pas prouvée.

9-9 : $\bar{x} = \bar{y} = 2,5$ $s_A = 1,29$ >> $s_B = 0,3$. Procédé B préférable (plus régulier).

9-10 : Si $\sigma_1 = \sigma_2$, $D = \frac{(n-1)S_1^{\,2}}{(m-1)S_2^{\,2}}$ suit $F_{n-1,m-1}$, etc...

10-6 :

classes	≤27	27-29	29-31	31-33	33-35	35-37	≥37
eff.th.	24,11	43,91	75,26	89,85	73,84	43,91	24,11
eff.obs.	27	38	79	84	75	53	19

loi normale (Test du khi-deux (ddl = 4) au niveau de risque 5%, 10%, ... $\alpha_{crit} \approx 0,3$)

10-7 : Test d'indépendance des caractères "occupation" et "mois". khi-deux (ddl = 1).

10-8 :

classes	≤30	30-60	60-90	≥90
eff.th.	45,12	24,76	13,59	16,53
eff.obs.	41	31	13	15

loi $\mathcal{E}(1/50)$ (Test du khi-deux (ddl = 3) au niveau de risque 5%, 10%, ... $\alpha_{crit} > 0,5$)

10-9 : indépendance rejetée au niveau de risque 5%, acceptée au niveau 2% (Test du khi-deux, ddl = 2).

10-10 : tirage au hasard de 50 copies à corriger par A, de 50 copies à corriger par B, test d'égalité de deux lois.

10-11 : Test des signes au risque 5% : Effet secondaire non démontré.