Contents

1 STATISTIQUE DESCRIPTIVE

1.1 Généralités et Exemples

Effectuer une étude statistique sur une population donnée, c'est tout d'abord relever les valeurs prises par une ou plusieurs variables au sein de cette population. Si une telle variable prend des valeurs numériques, on parlera de ***variable quantitative***, dans le cas contraire on parlera de ***variable qualitative***. Une variable quantitative sera dite ***discrète*** si ses valeurs possibles sont des valeurs numériques isolées (souvent des entiers) ; en revanche, si une variable quantitative prend ses valeurs dans tout un intervalle de $\mathbb{R}$, on parlera de variable quantitative ***continue***.

Prenons quelques **exemples simples** :

1. **Données brutes :** on relève la taille (en cm) d'une cinquantaine d'individus, les résultats sont regroupés ci-dessous :

158	*172*	*166*	*170*	*168*	*175*	*152*	*190*	*191*	*157*
163	*160*	*149*	*186*	*188*	*172*	*173*	*184*	*181*	*180*
172	*169*	*171*	*173*	*171*	*180*	*198*	*167*	*175*	*177*
170	*173*	*168*	*167*	*169*	*180*	*181*	*178*	*166*	*164*
160	*168*	*166*	*162*	*170*	*182*	*183*	*190*	*167*	*169*

   Ici, de toute évidence, nous avons affaire à une variable statistique *quantitative*.

2. **Données regroupées** ou **classées :** on relève le nombre annuel de visites de musées effectuées par 170 individus habitant Paris.

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
n_i	*12*	*34*	*43*	*39*	*28*	*10*	*4*

   Ici, la variable statistique X considérée (nombre annuel de visites) est de nature *quantitative* et *discrète*.
   Les nombres $n_0, n_1, \ldots, n_6$ que l'on a relevé dans cette population sont les *effectifs* de chacune des valeurs $x_0 = 0, x_1 = 1, \ldots, x_6 = 6$ de la variable X, leur somme $n = \sum_{i=0}^{6} x_i$ est aussi appelée effectif total, c'est la taille de l'ensemble de la population (ici $n = 170$).

3. On réalise une formule sanguine, c'est à dire que l'on observe au microscope une lame contenant des cellules sanguines, et on compte les cellules par type. On obtient alors les effectifs suivants :

Types de Cellules	*Effectifs*
Polynucléaires neutrophiles	***120***
Polynucléaires éosinophiles	***10***
Polynucléaires basophiles	***6***
Lymphocites	***54***
Monocytes	***10***

Les données sont encore présentées de façon regroupée, mais cette fois-ci elles sont de nature *qualitative*.

4. On s'intéresse au nombre d'enfants par famille dans une certaine ville, les observations sont effectuées sur un échantillon de 1500 ménages de la ville.

x_i	0	1	2	3	4	5	6	7	8
f_i	18%	24%	26,3%	16,5%	7%	4%	3%	0,7%	0,5%

Les valeurs possibles de la variable quantitative discrète X (nombre d'enfants dans la famille) sont les valeurs entières situées entre 0 et 8, et l'on indique dans ce tableau quelles sont les fréquences respectives $f_0, f_1, \ldots, f_8$ observées pour chacune de ces valeurs (où $f_i = \frac{effectif\ de\ x_i}{effectif\ total} = \frac{n_i}{n} = \frac{n_i}{1500}$).

5. Dans un grand lycée de la Région Parisienne, on considère la population constituée de tous les élèves de Terminale, en tout 500 individus, et l'on relève les valeurs prises par les quatre variables $W =$ Sexe (M, F), $X =$ Année de naissance $(1990, 1991, 1992, ...)$,
$Y =$ Série du Baccalauréat $(L, S, ES, ...)$ et $Z =$ Revenus annuels des parents $(KEuros/an)$.
Les résultats de cette enquête pourront être consignés en détail sous la forme de 500 quadruplets

$$(w_1, x_1, y_1, z_1),\ (w_2, x_2, y_2, z_2),\ (w_3, x_3, y_3, z_3),\ \ldots,\ (w_{500}, x_{500}, y_{500}, z_{500}).$$

Ici, manifestement : les variables W et Y sont *qualitatives*, tandis que X est une variable *quantitative discrète*. Z, quant à elle, pourra être considérée comme une variable *quantitative continue*.

6. Dans la Région Poitou-Charentes, on s'intéresse à une population de chênes de deux types (*"Pédonculé"* ou *"Pubescent"*), et l'on relève pour chacun d'eux la Pluviométrie Annuelle moyenne (en mm), la Température moyenne (en d^oC) et la nature du sol (acide, calcaire ou montagneux) de leur environnement.

7. Dans la Région PACA, on considère une population constituée de 500 entreprises, classées selon leurs types (*TPE*, *PME* ou *GE*). Pour chacune d'entre elles, on relève le nombre de salariés et le chiffre d'affaires annuel.
   Remarquons que dans les trois derniers exemples, on relève les valeurs de *plusieurs variables statistiques* simultanément, pour chacun des individus de la population considérée, et que cette population peut tout à fait être constituée d'individus qui ne sont pas des êtres humains (*chênes*, *entreprises*, ...).
   Voyons encore quelques exemples de présentations de données :

8. **Données regroupées en classes :** Le gestionnaire d'un Parc d'Attractions de la Région Parisienne décide de s'intéresser de près à la variable A : "Age d'un Visiteur", en considérant les 1000 prochains visiteurs de son parc et en relevanbt leurs âges respectifs $a_1, a_2, \ldots, a_{1000}$. On a affaire en l'occurence à une variable quantitative discrète, cependant les valeurs observées sont très diverses et plutôt nombreuses, il convient donc de les regrouper par classes, comme ci-dessous :

x_i	[0;10]	]10;15]	]15;20]	]20;30]	]30;40]	]40;60]	]60;80]
n_i	120	100	140	200	180	160	100

   Dans le cas d'une variable quantitative continue, ou encore dans celui d'une variable quantitative discrète prenant un grand nombre de valeurs, on sera souvent amené à ranger ces valeurs dans des intervalles, en général de type $]a;b]$, appelés ***classes statistiques***. La longueur $(b-a)$ de cette classe est appelée ***amplitude*** de la classe tandis que la valeur $\frac{a+b}{2}$ est appelée son centre ; le quotient de l'***effectif*** de la classe par son amplitude est appelé ***densité*** de la classe $d_i = \frac{n_i}{b-a}$.
   Lorsque l'on dispose de données quantitatives regroupées par classes, en l'absence d'informations supplémentaires, on sera souvent amené à utiliser une ***hypothèse de répartition uniforme*** : si $I \subset]a;b]$ est un sous-intervalle de $]a;b]$, l'effectif de I est proportionnel à sa longueur en tant que sous intervalle de la classe $]a;b]$. Ainsi, dans l'exemple ci-dessus, sous l'hypothèse de répartition uniforme on considérera qu'il y a 100 valeurs de la variable A situées dans l'intervalle $]20;25]$, et 100 autres valeurs situées dans l'intervalle $]25;30]$, ce qui correspond à une fréquence de 10% pour chacun de ces sous-intervalles de la classe $]20;30]$.
   Voyons une dernière situation où l'on est naturellement amené à présenter les données observées par classes, mais où cette fois ci les classes sont conçues différemment, le but étant d'obtenir une seule et même fréquence de 10% pour chacune des classes considérées.

9. Soit S la variable de salaire annuel net exprimée en kF (kilofrancs). Voici un tableau correspondant aux observations de l'INSEE pour l'année 1979 puis pour l'année 1983, la population étudiée étant celle des salariés à temps complet des secteurs privé et semi-public, à l'exclusion des personnels domestiques et de l'agriculture.

	1979	1983
1er Décile	*24,92*	*40,49*
2me Décile	*29,29*	*47,41*
3me Décile	*32,62*	*53,22*
4me Décile	*36,15*	*58,63*
5me Décile	*39,65*	*64,37*
6me Décile	*44,10*	*71,52*
7me Décile	*49,70*	*80,66*
8me Décile	*58,27*	*94,12*
9me Décile	*75,74*	*122,9*

Interprétation : en 1979, 10% des individus de la population étudiée avait un salaire annuel net situé entre le SMIC et $24,29\,kF$, pour 20% de ces individus le salaire annuel net était situé situé entre le SMIC et $29,29\,kF$, pour 90% de ces individus il était situé en-deçà de $75,74\,kF$.

1.2 Statistique à une dimension

Pour une variable statistique donnée, on pourra considérer les ***effectifs*** $n_1, n_2, \ldots$ associés aux différentes valeurs possibles ou encore aux différentes classes en présence.
La ***fréquence*** associée à la $i$ème valeur ou à la $i$ème classe est le quotient de l'effectif n_i par l'effectif total $n = \sum_i n_i$ de la population étudiée : $f_i = \frac{n_i}{n}$.
Une fréquence est donc un nombre situé entre 0 et 1, et la somme de toutes les fréquences vaut 1. Bien entendu, on pourra aussi exprimer cette répartition sur différentes valeurs ou classes en termes de ***pourcentages***, pour ce faire il suffit de multiplier les fréquences par 100.
Dans le cas d'une variable quantitative, on pourra toujours ranger les valeurs ou les classes dans leur ordre croissant. L'***effectif cumulé*** jusqu'à la valeur v_k vaut alors la somme des effectifs correspondants aux valeurs inférieures ou égales à v_k, et la ***fréquence cumulée*** s'obtient en divisant l'eff. cumulé par l'eff. total, ou encore en additionnant les fréquences associées aux valeurs $\leq v_k$.
D'une façon générale, la ***Statistique Descriptive*** s'attache à ***résumer*** et ***rendre intelligible*** les résultats d'une enquête statistique. Ces résumés ou explicitations se feront bien souvent au moyen de graphiques adaptés à la situation considérée.

1.2.1 Représentations Graphiques

- ***Diagrammes en bâtons :*** on trace au dessus de chaque valeur (ou encore au dessus de chaque centre de classe) un "bâton" dont la hauteur est proportionnelle à l'effectif

correspondant.
Ex. : tracer un diagramme en bâtons pour illustrer la situation statistique de l'exemple 2.

- ***Graphiques circulaires (camemberts) :*** on trace des portions de disque, et les angles au centre des portions sont proportionnels aux fréquences ou pourcentages en présence.
Ex. : tracer un graphique circulaire pour illustrer la situation statistique de l'exemple 3.

- ***Histogrammes :*** les intervalles des différentes classes sont reportés sur l'axe des abscisses, et servent de base à des *rectangles dont les aires sont proportionnelles aux effectifs, fréquences ou pourcentages* considérés. La hauteur de chaque rectangle correspond donc à la densité de la classe correspondante.
Ex. : tracer un histogramme pour illustrer la situation statistique de l'exemple 8.

- ***Fonction de répartition :*** des valeurs ou classes sont rangées par ordre croissant sur l'axe des abscisses, et à chaque valeur ou classe est associée sa fréquence cumulée ; la fonction obtenue croît de 0 à 1 et s'appelle *fonction de répartition* associée à la variable statistique quantitative étudiée.
Ex. : tracer une fonction de répartition pour illustrer la situation statistique de l'exemple 8, ou encore deux telles fonctions pour celle de l'exemple 9.

1.2.2 Paramètres de Position

Dans ce paragraphe ainsi que dans le paragraphe suivant (*Paramètres de Dispersion*), on s'intéresse à des paramètres attachés à une variable *quantitative* X, qui pourra être discrète ou continue. Soit donc $x_1, x_2, \ldots, x_n$ les valeurs prises par X, étudiée sur une population de taille n. Si X est discrète et ne prend qu'un nombre réduit de valeurs $v_1, v_2, \ldots, v_p$, on peut associer à chaque valeur v_i un effectif n_i et une fréquence $f_i = \frac{n_i}{n}$. Alternativement, on pourra être amené à considérer des effectifs et fréquences de classes.

1. ***Moyenne Arithmétique, ou Moyenne :*** c'est le nombre

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

Physiquement, cette définition correspond à un calcul de *Centre de Gravité*, mathématiquement on retrouve la notion de *Barycentre*.
Dans le cas d'une variable discrète ne prenant qu'un nombre réduit de valeurs, on aura aussi (cf. ex. 2 ou 4) :

$$\overline{x} = \sum_{j=1}^{p} f_j v_j = \left(f_1 v_1 + f_2 v_2 + \ldots + f_p v_p\right) = \frac{1}{n}\sum_{j=1}^{p} n_j v_j$$

Remarquons que

- La moyenne ne change pas si l'on remplace les effectifs par des effectifs proportionnels aux effectifs donnés :

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{p} n_j v_j = \frac{1}{10n}\sum_{j=1}^{p} (10n_j) v_j = \frac{1}{n/10}\sum_{j=1}^{p} (n_j/10) v_j$$

- On pourra commencer par effectuer des moyennes $\overline{x}_1, \overline{x}_2, \overline{x}_3$ sur des sous-populations, de tailles respectives N_1, N_2, N_3 avec $N_1 + N_2 + N_3 = n$, puis effectuer la moyenne de ces moyennes partielles en respectant les différentes tailles de sous-populations :

$$\overline{x} = \frac{N_1\overline{x}_1 + N_2\overline{x}_2 + N_3\overline{x}_3}{N_1 + N_2 + N_3} = \frac{N_1\overline{x}_1 + N_2\overline{x}_2 + N_3\overline{x}_3}{n}$$

- Lorsque les valeurs sont rangées par classes, on obtiendra une approximation de la moyenne en remplaçant chacune de ces valeurs par le centre de sa classe :

$$\overline{x} \approx \sum_{j=1}^{p} f_j c_j = \left(f_1 c_1 + f_2 c_2 + \ldots + f_p c_p\right) = \frac{1}{n}\sum_{j=1}^{p} n_j c_j$$

- Il est bien évident que la valeur moyenne $\overline{x}$ de la variable statistique x sera toujours située *entre la plus petite et la plus grande des valeurs observées.* Ceci résulte de la monotonie du passage à la moyenne : si l'on observe simultanément deux variables quantitatives x et y, et que pour chaque individu i de la population les observations x_i et y_i sont telles que $x_i \leq y_i$, on aura $\overline{x} \leq \overline{y}$.
- En revanche, bien souvent la valeur moyenne $\overline{x}$ n'apparaît pas parmi les valeurs effectivement observées au cours de l'enquête (cf. exemple 4 : le nombre moyen d'enfants par famille s'élève à $\overline{x} = 2,01$).

2. ***Médiane :***
   c'est une valeur m_X qui est telle que : pour 50% de la population, la valeur de X est située sous m_X, et pour les 50% restants elle est située au dessus.
   A supposer que le nombre n d'observations de la variable X est impair, on pourra ordonner ces observations par ordre croissant puis considérer la $\left(\frac{n+1}{2}\right)$ème valeur comme valeur médiane. Dans le cas d'un nombre pair d'observartions, il conviendra d'effectuer la demi-somme des $\left(\frac{n}{2}\right)$ème et $\left(\frac{n}{2}+1\right)$ème valeurs pour calculer m_X.

3. ***Quantiles :***
   on associe à chaque niveau α situé entre 0 et 1 (ou entre 0% et 100%) un ***quantile*** q_α, valeur telle que *pour une proportion α de la population la valeur de X est située sous q_α.*
   On définit en particulier les ***quartiles*** Q_1, Q_2, Q_3, Q_4 comme étant les quantiles associés respectivement aux niveaux $25\%, 50\%, 75\%$ et 100%, et les ***déciles*** $D_1, D_2, D_3, \ldots, D_{10}$ comme étant les quantiles associés respectivement aux niveaux $10\%, 20\%, 30\%, \ldots, 100\%$ (cf ex. 9).
   La médiane m_X coïncide donc avec le deuxième quartile Q_2, ou encore avec le cinquième décile D_5.

4. ***Mode :*** c'est tout simplement la valeur la plus fréquemment prise, ou encore la classe de valeurs ayant un effectif maximal.

N.B. : un très grand écart entre moyenne et médiane est possible, c'est un indice de grande ***dispersion*** des valeurs, ce que l'on pourra observer par exemple dans une étude de salaires annuels !

1.2.3 Paramètres de Dispersion

1. ***Variance :*** il s'agit d'un paramètre $V(X) \geq 0$ qui admet deux définitions équivalentes,

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{p} n_j(v_j - \overline{x})^2 = \sum_{j=1}^{p} f_j(v_j - \overline{x})^2$$

   (variance comme ***moyenne des carrés des écarts à la moyenne***), ou encore

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

(variance comme ***différence entre moyenne des carrés et carré de la moyenne***). Remarquons que

- La variance est sensible aux changements d'échelle :

$$V(X/c) = V(X)/c^2,$$

(penser par ex. à des statures mesurées en m ou en cm), en revanche elle est insensible aux translations :

$$V(X + c) = V(X)$$

- Pour obtenir un paramètre de dispersion mesuré suivant les mêmes unités que X, on définit l'***écart-type*** σ comme racine carrée de la variance :

$$\sigma_X = \sqrt{V(X)} = \sqrt{\overline{x^2} - \overline{x}^2}$$

- Lorsque les valeurs sont rangées par classes, on obtiendra une approximation de la variance en remplaçant chacune de ces valeurs par le centre de sa classe.

2. ***Etendue :*** c'est juste la différence entre les plus grande et plus petite valeurs prises par une variable quantitative. (Bien entendu, la moyenne est toujours située entre ces plus petite et plus grande valeurs).

3. ***Intervalle interdécile, écart interquartile :***
L'***intervalle interdécile*** est l'intervalle $[D_1; D_9]$, tandis que l'***écart interquartile*** vaut $(Q_3 - Q_1)$.
Les premiers et neuvièmes déciles D_1 et D_9 sont parfois utilisés conjointement aux premier et troisième quartiles et à la médiane pour tracer une ***Boîte de Dispersion***, parfois appelée ***"Boîte à Moustaches"*** (en France).

1.2.4 Exercices d'Application

I Reprendre les exemples 1 à 9 de cette section pour calculer, à chaque fois que cela est possible, les valeurs moyennes, médianes et variances des variables statistiques étudiées.

II On désire calculer la moyenne d'une promotion d'étudiants à une épreuve de mathématiques, et l'on sait que

- le groupe A (40 étudiants) a une moyenne de $10,9$
- le groupe B (45 étudiants) a une moyenne de $9,8$
- le groupe C (52 étudiants) a une moyenne de 11
- le groupe D (32 étudiants) a une moyenne de $9,5$

a) Calculer (en expliquant votre démarche) la moyenne de la promotion.

b) Le correcteur du groupe D a reçu 8 copies en retard, la moyenne de ces 8 copies supplémentaires s'élève à $4,5$.
Quelle est la nouvelle moyenne du groupe D ? De la promotion entière ??

c) On calcule l'écart-type de ces notes au sein de chaque groupe, les résultats obtenus sont les suivants :

- le groupe A (40 étudiants) a un écart-type de $2,8$
- le groupe B (45 étudiants) a un écart-type de $3,3$
- le groupe C (52 étudiants) a un écart-type de $2,6$
- le groupe D (32 étudiants) a un écart-type de $3,5$

Calculer (en expliquant votre démarche) l'écart-type de la promotion.

III On s'intéresse à la variable d de durée de vie d'un certain type d'ampoule. On a mesuré la durée de vie (en semaines) de 500 ampoules de ce type et obtenu, dans un premier temps, le Tableau 1 ci-dessous, puis le Tableau 2, qui diffère du précédent par le choix des classes utilisées :

Tableau 1 :

[0;40]	*10*
]40;80]	*50*
]80;120]	*220*
]120;160]	*130*
]160;200]	*50*
]200;240]	*30*
]240;280]	*5*
]280;320]	*3*
]320;360]	*1*
]360;400]	*1*

Tableau 2 :

[0;40]	*10*
]40;60]	*15*
]60;80]	*35*
]80;90]	*30*
]90;100]	*50*
]100;110]	*60*
]110;120]	*80*
]120;130]	*50*
]130;140]	*30*
]140;160]	*50*
]160;180]	*30*
]180;200]	*20*
]200;240]	*30*
]240;400]	*10*

a) Tracer les histogrammes et fonctions de répartition obtenus à partir de chacun des deux tableaux.
Quel lien observez-vous entre histogrammes et fonctions de répartition ?

b) Calculer une valeur approchée de la valeur moyenne $\hat{d}$ en vous aidant du Tableau 1, puis en vous aidant du Tableau 2.

c) Calculer une valeur approchée de la variance puis de l'écart-type de la variable d en vous aidant du Tableau 1, puis en vous aidant du Tableau 2.

IV On a relevé la portée (en mètres) de 5 des plus grands ponts suspendus du monde :

Pont	Seto	Innoshima	Mackinac	Oakland	Narrows
Pays	***Japon***	***Japon***	***USA***	***USA***	***USA***
Portée	***1100***	***770***	***1158***	***704***	***1298***

Calculer la valeur moyenne ainsi que l'écart-type de cette distribution statistique en vous aidant du fait que

$$\sum_{i=1}^{5} x_i = 5\,030 \text{ et } \sum_{i=1}^{5} x_i^2 = 5\,324\,284$$

V On a relevé les taux d'hémoglobine dans le sang de 60 adultes en bonne santé ; les valeurs de cette série statistique (en g/L) sont notées ci-dessous dans leur ordre non décroissant, les taux mesurés sur des femmes étant reportés en caractères gras :
105 ; **110** ; **112** ; **112** ; **118** ; **119** ; **120** ; **120** ; **125** ; **126** ; **127** ; **128** ; **130** ; **132** ; **133** ; **134** ; **135** ; **138** ; **138** ; **138** ; **138** ; 141 ; **142** ; 144 ; **145** ; 146 ; **148** ; **148** ; 148 ; 149 ; **150** ; 150 ; 150 ; **151** ; 151 ; 153 ; 153 ; 153 ; **154** ; **154** ; 154 ; 155 ; 156 ; 156 ; **158** ; 160 ; 160 ; 160 ; 163 ; 164 ; 164 ; 165 ; 166 ; 168 ; 168 ; 170 ; 172 ; 172 ; 176 ; 179.
On calcule en outre les sommes suivantes :

$$\text{pour les hommes, } \sum_{i=1}^{30} x_i^h = 4'766\,(g/L), \quad \sum_{i=1}^{30} (x_i^h)^2 = 759'954\,(g/L)^2,$$

$$\text{pour les femmes, } \sum_{i=1}^{30} x_i^f = 3'988\,(g/L), \quad \sum_{i=1}^{30} (x_i^f)^2 = 536'176\,(g/L)^2.$$

(a) Considérons le regroupement suivant les classes

$$]104;114]\,;\ ]114;124]\,;\ ]124;134]\,;\ ]134;144]\,;\ ]144;154]\,;\ ]154;164]\,;\ ]164;174]\,;\ ]174;184]\,.$$

Pour chacune des deux séries (hommes, femmes), déterminer les effectif et fréquence de chaque classe.

(b) Donner une représentation graphique appropriée de chacune de ces deux séries rangées par classes.

(c) Calculer les moyennes $\overline{x}, \overline{x^h}, \overline{x^f}$ de la série initiale puis des séries masculine et féminine.

(d) Calculer les nouvelles moyennes obtenues en plaçant chaque valeur dans une classe puis en remplaçant cette valeur par le milieu de la classe.

(e) Donner les médianes m_X, m_{X^h}, m_{X^f} des trois séries considérées.

(f) Calculer l'écart interquartiles puis donner l'intervalle interdéciles pour chacune des trois séries. Dessiner les boîtes de dispersion ("boîtes à moustache") associées à ces séries.

(g) Calculer les variances et écarts-types de ces trois séries.

(h) Quelles variances et écarts-types obtient-on en plaçant chaque valeur dans une classe puis en remplaçant cette valeur par le milieu de la classe ?

1.3 Statistique à deux dimensions

Dans les divers exemples considérés jusqu'ici, on a relevé simultanément *plusieurs valeurs* de variables statistiques pour chaque individu d'une population donnée (Sexe *et* Année de naissance *et* Série du Bac *et* Revenus annuels, ou encore Sexe *et* Taux d'Hémoglobine).

1.3.1 Distributions à deux dimensions

1. Déterminer une ***distribution statistique à deux dimensions*** pour le couple de variables (X, Y) revient à spécifier

   - les valeurs possibles $v_1, v_2, \ldots, v_p$ pour la variable X
   - les valeurs possibles $w_1, w_2, \ldots, w_q$ pour la variable Y
   - les effectifs $n_{i,j}$ correspondant aux différentes éventualités du type ($X = v_i$ et $Y = w_j$).

   Si l'effectif total est de taille n, la fréquence correspondant à l'observation ($X = v_i$ et $Y = w_j$) vaut $f_{i,j} = \frac{n_{i,j}}{n}$. Ces données sont souvent présentées dans un tableau à double entrée.
   Exemple : Statures et Pointures à l'intérieur d'un groupe.

2. A partir d'une distribution statistique à deux dimensions pour le couple de variables (X, Y), on pourra retrouver les ***distributions marginales*** relatives à la seule variable X ou à la seule variable Y :

   - $(X = v_i)$ a pour effectif $n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j}$ et pour fréquence $f_{i,\cdot} = \sum_{j=1}^{q} f_{i,j}$
   - $(Y = w_j)$ a pour effectif $n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j}$ et pour fréquence $f_{\cdot,j} = \sum_{i=1}^{p} f_{i,j}$
   - Ces ***effectifs marginaux*** ou ***fréquences marginales*** s'obtiennent donc par des *additions suivant les lignes* ou *suivant les colonnes* dans le tableau à double entrée, additions dont les résultats sont reportés en marge du tableau.

3. La ***distribution conditionnelle*** de Y pour $X = v_i$ s'obtient en consignant les valeurs prises par Y seulement pour les $n_{i,\cdot}$ individus satisfaisant $X = v_i$.
   Similairement, la ***distribution conditionnelle*** de X pour $Y = w_j$ s'obtient en consignant les valeurs prises par X seulement pour les $n_{\cdot,j}$ individus satisfaisant $Y = w_j$.

4. Les variables statistiques X et Y seront dites ***indépendantes*** dès lors que les fréquences conjointes $f_{i,j}$ sont *systématiquement égales* au produit $f_{i,\cdot} \times f_{\cdot,j}$ des fréquences marginales correspondantes :
   $$\forall i \in [\![1; p]\!],\ \forall j \in [\![1; q]\!],\ f_{i,j} = f_{i,\cdot} \times f_{\cdot,j},$$
   ou encore
   $$\forall i \in [\![1; p]\!],\ \forall j \in [\![1; q]\!],\ n_{i,j} = \frac{n_{i,\cdot} \times n_{\cdot,j}}{n}$$
   Pour qu'il y ait indépendance statistique entre X et Y, il faut donc qu'une telle égalité soit valable *pour chacune des cases du tableau !* La mise en défaut de cette égalité pour une

seule case suffit à prouver que X et Y ne sont pas indépendantes.
Remarquons en outre que

- l'indépendance statistique de X et Y correspond au fait que les lignes du tableau à double entrée sont proportionelles entre elles, ainsi que les colonnes.
- Cette indépendance statistique de X et Y correspond aussi à une *indépendance de Y par rapport à X*, au sens où les fréquences conditionnelles de Y pour $X = v_i$ ne dépendent pas de la valeur v_i, et à une *indépendance de X par rapport à Y*, au sens où les fréquences conditionnelles de X pour $y = w_j$ ne dépendent pas de la valeur w_j.

1.3.2 Paramètres associés à une série statistique double

Considérons à nouveau une série statistique double correspondant aux valeurs prises *conjointement* par deux variables X et Y. On pourra appliquer les définitions précédentes aux distributions marginales associées à X et à Y, en sorte que

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i,.}v_i \text{ et } \overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.,j}w_j$$

Le point de coordonnées $(\overline{x}, \overline{y})$ est appelé ***point moyen*** associé à cette série statistique double. Les variances marginales $V(X)$ et $V(Y)$ pourront aussi être calculées comme

$$V(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,.}(v_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{p} n_{i,.}v_i^2\right) - \overline{x}^2$$

et

$$V(Y) = \sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.,j}(w_j - \overline{y})^2 = \frac{1}{n}\left(\sum_{j=1}^{q} n_{.,j}w_j^2\right) - \overline{y}^2$$

Venons-en aux définitions réellement nouvelles de ***Covariance*** et de ***Corrélation*** :

1. La ***Covariance*** $\mathrm{Cov}(X, Y)$ est définie comme ***moyenne des produits des écarts à la moyenne***, ou encore comme ***différence entre la moyenne des produits et le produit des moyennes*** :

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{1\leq i\leq p, 1\leq j\leq q} n_{i,j}(v_i - \overline{x})(w_j - \overline{y}) = \frac{1}{n}\left(\sum_{1\leq i\leq p, 1\leq j\leq q} n_{i,j}v_i w_j\right) - \overline{x}\times\overline{y}$$

   Il est important de noter que

   - La covariance généralise la variance, au sens où $\mathrm{Cov}(X, X) = V(X)$.
   - $\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y)$

- $|\text{Cov}(X,Y)| \leq \sigma_X \sigma_Y$
- Si les variables statistiques X et Y sont indépendantes, $\text{Cov}(X,Y) = 0$.
  (Attention cependant, la réciproque est ici fausse ! On pourra avoir $\text{Cov}(X,Y) = 0$ sans pour autant que les variables statistiques X et Y soient indépendantes).

2. Pour obtenir un paramètre conjoint insensible aux changements d'échelle, on définit le ***Coefficient de Corrélation Linéaire*** $r = r(X,Y)$ comme

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Notons que

- $r(aX+b, cY+d) = r(X,Y)$
- On aura toujours $-1 \leq r(X,Y) \leq 1$.
- Si les variables statistiques X et Y sont indépendantes, $r(X,Y) = 0$.
- L'égalité $r(X,Y) = 1$ ne se produira que si tous les points obtenus en relevant les valeurs prises simultanément par X et Y sont situés sur une même droite de pente positive. $r(X,Y) = -1$ correspond à la situation où ces points sont alignés sur une même droite de pente négative.

1.3.3 Modèles d'ajustement, méthode des moindres carrés

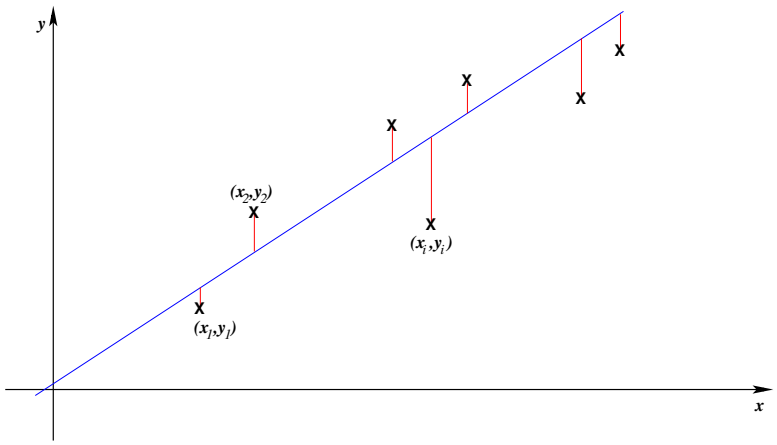
Considérons à nouveau un ***nuage de points*** $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ correspondant aux valeurs prises conjointement par deux variables statistiques X et Y au sein d'une population de taille n.
L'allure de ce nuage de points et des considérations sur le phénomène étudié peuvent suggérer une relation fonctionnelle entre les grandeurs x et y, comme par exemple

$$y = ax + b, \quad y = ax^b \text{ ou encore } y = a \ln x + b$$

Les modèles d'ajustement ci-dessus comportent chacun deux paramètres a et b.
Après avoir choisi un modèle, on utilisera une certaine distance entre les courbes du type choisi et la courbe expérimentale afin de déterminer des valeurs optimales des paramètres a et b, valeurs qui permettent de minimiser la distance entre la courbe théorique et la courbe expérimentale.
Quand les points de la courbe expérimentale sont à peu près alignés, on pourra retenir le modèle

$$y = ax + b \text{ (ajustement affine).}$$

Dans la ***méthode des moindres carrés***, il s'agit alors de régler les paramètres a et b afin de minimiser la *somme des carrés des écarts verticaux* entre les points expérimentaux et la droite théorique :



On cherche donc à minimiser

$$S(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Il s'avère que les valeurs optimales de a et b donnent une droite *passant par le point moyen* de coordonnées $(\overline{x},\overline{y})$ et *ayant pour pente le quotient de la covariance* $Cov(X,Y)$ *par la variance* $V(X)$. Ces valeurs optimales $\hat{a}$ et $\hat{b}$ sont donc déterminées par les relations

$$\begin{cases} \hat{a} = \frac{\mathrm{Cov}(X,Y)}{V(X)} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y}}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2} \\ \hat{a}\overline{x} + \hat{b} = \overline{y} \end{cases}$$

Discussion sur la Régression Linéaire :

(a) Remarquons que dans une telle recherche de ***Droite de Régression de*** Y ***par rapport à*** X, les deux variables en présence ne jouent pas le même rôle :
Y est la variable "à expliquer", tandis que X constitue la variable *potentiellement* explicative - il ne faut d'ailleurs pas se hâter d'expliquer les variations de Y à partir de celles de X ... Il y a parfois une troisième variable qui se cache entre les variables X et Y !

(b) Dans le but d'évaluer la qualité d'une régression linéaire, on pourra commencer par observer que

$$V(Y) = V(\hat{a}X + \hat{b}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2,$$

autrement dit :

Variance de Y = Variance expliquée par l'ajustement affine + Variance résiduelle

On a ensuite

$$\frac{\text{Variance expliquée}}{\text{Variance totale}} = \frac{V(\hat{a}X + \hat{b})}{V(Y)} = \hat{a}^2 \frac{V(X)}{V(Y)} = \frac{\text{Cov}(X,Y)^2}{V(X)V(Y)} = r^2$$

Il conviendra donc d'utiliser la droite de régression linéaire seulement si le coefficient de corrélation $r = r(X,Y)$ est suffisamment élevé, mettons si $|r| > \frac{1}{\sqrt{2}} \approx 0,71$, seuil à partir duquel les variations de $(\hat{a}X + \hat{b})$ contribuent à plus de 50% de celles de Y.

1.3.4 Exercices d'Application

I Dans une population constituée de ménages ayant des enfants, on procède à une étude simultanée de deux variables statistiques quantitatives : le nombre d'enfants X, et l'âge Y du premier enfant.
Les effectifs obtenus sont reportés dans le tableau à double entrée ci-dessous :

X \ Y	de 0 à 4	de 5 à 9	de 10 à 14	de 15 à 19	de 20 à 24
1	30	28	35	43	21
2	26	35	32	31	27
3	20	29	26	23	18
4	2	15	18	16	19
5	0	3	4	5	10

Déterminer les distributions marginales de X et de Y.
Ces variables statistiques sont-elles indépendantes ?

II On a demandé à vingt adolescents âgés de 12 à 16 ans combien d'argent de poche ils recevaient chaque semaine (en Euros). Le tableau ci-dessous récapitule leurs réponses :

Individu i	1	2	3	4	5	6	7	8	9	10
Age x_i	12	12	15	14	16	14	12	13	15	13
Montant y_i	4	5	12	11	12	8	5	8	6	4

Individu i	11	12	13	14	15	16	17	18	19	20
Age x_i	13	15	16	14	14	16	15	12	13	16
Montant y_i	7	11	13	6	7	10	10	3	6	12

a) Calculer les valeurs moyennes $\overline{x}, \overline{y}$ associées à ces deux variables statistiques, puis leurs variances et leurs écarts-types.

b) Calculer la covariance de ces deux variables, puis le coefficient de corrélation linéaire r et le coefficient de détermination r^2. Que peut-on en conclure ?

c) Tracer le nuage de points correspondant à ces données, puis la droite de régression linéaire associée, dont on donnera l'équation cartésienne.

d) Dans ces conditions, combien d'argent un adolescent peut-il espérer recevoir hebdomadairement à l'âge de 17 ans ?

III Dans une population donnée, on désire voir si la tension artérielle Y (mesurée en $cmHg$) est fortement corrélée avec l'âge X (mesuré en années). Après mesures et calculs, on obtient les moyennes

$$\overline{x} = 35, \quad \overline{y} = 13,5$$

ainsi que les variances et covariance

$$V(X) = 64, \quad V(Y) = 4, \quad \mathrm{Cov}(X,Y) = 10.$$

Evaluer le coefficient de corrélation $r = r(X,Y)$ puis commenter la situation.

IV On admet l'existence d'une relation affine entre le logarithme de la dose de vitamine contenue dans le milieu de culture d'une colonie bactérienne et le diamètre de cette culture.
Au vu des résultats collectés dans le tableau ci-dessous :

Dose en µg	***10***	***20***	***40***
Diam. (mm)	***2 ; 3 ; 2***	***3 ; 5 ; 4***	***6 ; 7 ; 6***

quelle est la meilleure estimation de la dose de vitamine contenue dans un milieu où la colonie bactérienne aurait un diamètre de $3mm$?

V Le tableau ci-dessous fournit la série des productions et des prix de pommes à cidre entre 1888 et 1916 :

Année	Prix	Produc.
1888	8,12	12,7
1889	9,71	4,6
1890	9,77	9,6
1891	10,19	8,2
1892	6,69	15,9
1893	3,14	38,8
1894	6,13	16,9
1895	4,18	28,2
1896	7,33	11,4
1897	10,27	7,9
1898	9,66	10,6
1899	5,76	22,5
1900	3,5	37,5
1901	6,85	12,6
1902	11,79	8,6
1903	12,68	5,4
1904	2,91	62,6
1905	12,2	4,6
1906	5,04	26,1
1907	13,48	4,2
1908	6,18	22,4
1909	8,00	12,7
1910	9,02	13,4
1911	6,41	30,8
1912	6,54	23,9
1913	4,06	51,2
1914	4,19	24,2
1915	3,99	37,2
1916	11,45	8,8

a) Représenter graphiquement le nuage de points correspondant à ces données, puis rechercher l'équation $y = \hat{a}x + \hat{b}$ de la droite de régression linéaire correspondante.

b) Chercher un changement de variables permettant de réaliser un ajustement du type $y = \alpha x^{\beta}$, puis donner les valeurs optimales des paramètres α et β dans un tel modèle.

c) Comparer les deux ajustements obtenus en a) et en b).
Commentaire, interprétation économique ?

VI Au cours d'un Test d'Effort dans une Unité de Cardiologie, on fait varier l'intensité X (en kJ/mn) du travail fourni par un patient, et on relève systématiquement la fréquence cardiaque Y (nombre de battements/mn) correspondante. Les résultats sont consignés dans le tableau ci-dessous :

x_i	9,6	12,8	18,4	31,2	36,8	47,2	49,6	56,8
y_i	70	86	90	104	120	128	144	154

a) Représenter ces résultats par un nuage de points.

b) Calculer le coefficient de corrélation $r = r(X, Y)$.

c) Déterminer la droite de régression de Y par rapport à X, puis décomposer la variance de Y comme somme de la variance expliquée par l'ajustement affine et de la variance résiduelle.

d) Estimer la fréquence cardiaque du patient lorsque l'intensité du travail fourni est de $30kJ/mn$, puis lorsqu'elle est de $75kJ/mn$. Commentaire ?

VII Une étude théorique de l'évolution d'une population en exctinction conduit à penser que le nombre d'individus N de cette population varie avec le temps t (mesuré en mois) suivant une loi exponentielle du type $N(t) = ae^{-kt}$, où a et k sont des constantes strictement positives.
On cherche à déterminer expérimentalement la valeur de la constante k. Pour ce faire, on observe pendant 8 mois un échantillon composé de 200 individus, notant à la fin de chaque mois le nombre de survivants :

t	1	2	3	4	5	6	7	8
$N(t)$	180	154	140	120	112	97	84	76

a) En utilisant un ajustement affine, déterminer une valeur approchée de la constante k.

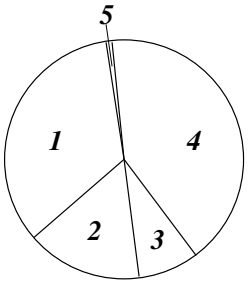
b) Estimer le nombre de survivants dans cette population à l'issue de l'année en cours, puis à la fin de l'année suivante.

1.4 Questionnaires à choix multiples

I Avant une élection entre cinq candidats, on a demandé à un échantillon d'électeurs pour quel candidat ils avaient l'intention de voter et obtenu les résultats suivants :

- pour le candidat A : 190 intentions de votes.
- pour le candidat B : 10 intentions de votes.
- pour le candidat C : 380 intentions de votes.
- pour le candidat D : 70 intentions de votes.
- pour le candidat E : 300 intentions de votes.

On représente ces résultats au moyen du diagramme circulaire ci-dessous :



QCM1 : Quelle(s) est(sont) alors la(les) proposition(s) exactes ?

A Le candidat A a obtenu 20% des intentions de votes.

B Le candidat A a obtenu 25% des intentions de votes.

C Les résultats du candidat A correspondent au secteur no.1 du diagramme.

D Les résultats du candidat A correspondent au secteur no.2 du diagramme.

E Les résultats du candidat A correspondent au secteur no.3 du diagramme.

II Dans un pays lointain, des conscrits ont été pesés au début de leurs service militaire, les résultats obtenus ont ensuite été résumés dans le tableau suivant :

POIDS (Kg)	***]50;55]***	***]55;60]***	***]60;65]***	***]65;70]***	***]70;75]***	***]75;80]***	***]80;85]***	***]85;90]***	***]90;95]***	***]95;100]***
Nb d'individus	***5***	***10***	***17***	***25***	***30***	***27***	***18***	***9***	***6***	***3***

QCM2 : Si l'on trace un histogramme normalisé, où les rectangles ont des aires correspondant aux fréquences des différentes classes de poids, quelle est la hauteur du rectangle dont la base est la classe $]70Kg; 75Kg]$?

A $0,02$

B $0,03$

C $0,04$

D $0,2$

E $0,3$

QCM3 : Toujours dans un histogramme normalisé, si l'on trace un rectangle de base $]85Kg; 100Kg]$, quelle sera sa hauteur ?

A $0,002$

B $0,008$

C $0,012$

D $0,12$

E $0,18$

III Sur un échantillon de 13 patients atteints de troubles bipolaires, on a noté l'âge de début de leur maladie et obtenu les données suivantes :

PATIENT	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Age début maladie	***17***	***28***	***40***	***31***	***33***	***46***	***25***	***52***	***49***	***37***	***21***	***15***	***51***

On note :

- $\min(a)$ le minimum des 13 valeurs d'âge observées
- $\max(a)$ le maximum des 13 valeurs d'âge observées
- $\text{Etendue}(a) = \max(a) - \min(a)$

QCM4 : Quelle(s) est(sont) la(les) expression(s) exacte(s) ?

A $\min(a) = 17$

B $\min(a) = 21$

C $\max(a) = 51$

D $\max(a) = 52$

E $\text{Etendue}(a) = 35$

On note :

- $m(a)$ la moyenne statistique de l'âge de début de la maladie
- $s(a)^2$ sa variance statistique
- $s(a)$ son écart-type statistique.

QCM5 : Quelle(s) est(sont) la(les) expression(s) exacte(s) (à $0,1$ près) ?

A $m(a) = 34,2$

B $m(a) = 37,5$

C $s(a)^2 = 152,1$

D $s(a)^2 = 164,4$

E $s(a) = 12,3$

On note :

- $Q_1(a)$ le premier quartile de cet échantillon d'observations
- $Q_3(a)$ le troisième quartile de cet échantillon d'observations
- médiane(a) sa médiane.

QCM6 : Quelle(s) est(sont) la(les) expression(s) exacte(s) ?

A $Q_1(a) = 21$

B $Q_1(a) = 25$

C $Q_3(a) = 40$

D $Q_3(a) = 46$

E médiane$(a) = 32$

QCM7 : Si l'on devait remplacer la valeur 21 apparaissant dans cet échantillon par 101, quelles statistiques s'en trouveraient modifiées ?

A $m(a)$

B $s(a)$

C $Q_1(a)$

D $Q_3(a)$

E médiane(a)

QCM8 : Si maintenant c'est la valeur 51 et non plus la valeur 21 qui est remplacée par la valeur 101, quelles statistiques s'en trouvent modifiées ?

A $m(a)$

B $s(a)$

C $Q_1(a)$

D $Q_3(a)$

E médiane(a)

IV Sur un échantillon de 100 individus, on a mesuré deux variables x, y puis calculé les statistiques suivantes :

$$m_x = 75,\ m_y = 70,\ s_x = 10,\ s_y = 8,\ r(x,y) = 0,9,$$

où m, s et r désignent respectivement des moyennes, écarts-type et coefficients de corrélation statistiques.
A partir de ces indications, on cherche à trouver $\Sigma_{x^2} = \sum_{i=1}^{100} x_i^2$ et $\Sigma_{y^2} = \sum_{i=1}^{100} y_i^2$

QCM9 : Quelle(s) est(sont) la(les) expression(s) exacte(s) ?

A $\Sigma_{x^2} = 563'490$

B $\Sigma_{x^2} = 572'400$

C $\Sigma_{y^2} = 483'664$

D $\Sigma_{y^2} = 490'792$

E $\Sigma_{y^2} = 496'336$

On s'intéresse maintenant à la variable $d = x - y$, et l'on cherche à calculer sa moyenne et son écart-type statistiques m_d et s_d. On note aussi :

- $\text{Cov}(x, y)$ la covariance des variables x et y
- $\Sigma_{xy} = \sum_{i=1}^{100} x_i y_i$

QCM10 : Quelle(s) est(sont) la(les) expression(s) exacte(s) ?

A $\text{Cov}(x, y) = 72$

B $\text{Cov}(x, y) = 5'760$

C $\Sigma_{xy} = 496'336$

D $\Sigma_{xy} = 532'128$

E $\Sigma_{xy} = 1'095'240$

QCM11 : Quelle(s) est(sont) la(les) expression(s) exacte(s) à $0,01$ près ?

A $m_d = 5$

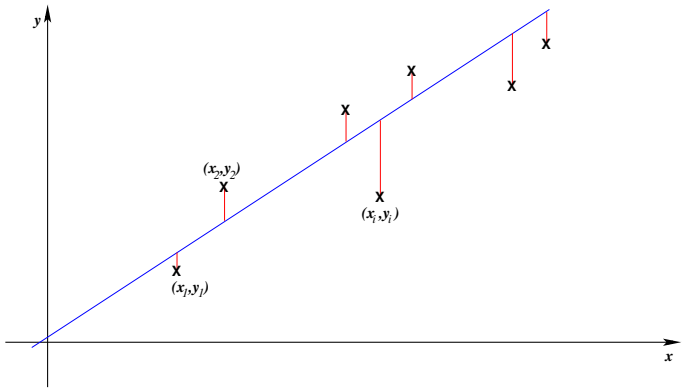
B $m_d = 8,45$

C $s_d = 4,47$

D $s_d = 5,12$

E $s_d = 8,45$

V On considère deux variables statistiques x et y dont les valeurs forment un nuage de points esquissé ci-dessous.



QCM12 : Au vu de ce nuage, on peut affirmer que

A Les variables x et y sont indépendantes.

B Les variables x et y sont positivement corrélées.

C Les variables x et y sont négativement corrélées.

D La valeur de x ne me renseigne aucunement sur la valeur de y.

E Si y prend une basse valeur, on tend à estimer que la valeur de x est élevée.

VI On considère deux variables statistiques z et t pour lesquelles les calculs des paramètres usuels ont donné

$$\overline{z} = 22,8,\ \overline{t} = 3,7,\ \mathrm{Cov}(z;t) = -18,\ \mathrm{Var}(z) = 16,\ \mathrm{Var}(t) = 25$$

Au vu de ce résultats, on peut affirmer que
QCM13 :

A Les variables z et t sont indépendantes.

B La médiane de z vaut $22,8$.

C Le coeff. de corrélation linéaire $r(z;t)$ vaut $-0,9$.

D La valeur de z ne me renseigne aucunement sur la valeur de t.

E Si z prend une basse valeur, on tend à estimer que la valeur de t est basse.

mais aussi que
QCM14 :

A Dans le modèle de régression linéaire $t = \hat{a}z + \hat{b}$, la pente $\hat{a}$ vaut $-\frac{9}{8}$.

B Dans le modèle de régression linéaire $t = \hat{a}z + \hat{b}$, la pente $\hat{a}$ vaut $-0,9$.

C Dans le modèle de régression linéaire $t = \hat{a}z + \hat{b}$, $\hat{b}$ vaut $-\frac{9}{8}$.

D Dans le modèle de régression linéaire $t = \hat{a}z + \hat{b}$, $\hat{b}$ vaut $-3,7$.

E Ce modèle de régression linéaire ne doit pas être utilisé car ici $r^2 < \frac{1}{2}$.

VII On souhaite évaluer la quantité journalière moyenne de nourriture nécessaire à un enfant pendant les premiers mois de la vie, en fonction de son poids. Pour cela, on observe une vingtaine d'enfants sains pendant une semaine, et on note leurs poids x (en grammes) ainsi que les quantités y de lait qu'ils ont consommées (en mL).
On obtient les moyennes $\overline{x} = 5200$, $\overline{y} = 770$, ainsi que les écarts-types $s_X = 500, s_Y = 80$ et la covariance $Cov(X,Y) = 25000$.
QCM15 : A partir de ces données empiriques, quelle est l'équation cartésienne de la droite de régression linéaire de y en fonction de x ?

A $y = 0,1\,x + 250$

B $y = 250\,x + 0,1$

C $y = 0,25\,x + 100$

D $y = 0,125\,x$

E $y = 0,1\,x - 150$

Un nourrisson âgé de 5 mois et pesant 7 Kg a consommé, durant la journée, 3 biberons de 270mL de lait.
QCM16 : Quelle quantité de lait doit-on prévoir pour son dernier biberon ?

A 140 mL.

B 440 mL.

C 950 mL.

D La quantité journalière est déjà dépassée.

E Informations insuffisantes pour donner une réponse précise.

2 CALCUL des PROBABILITES

Dans ce deuxième chapitre, le point de vue est différent de celui pris en Statistique : on se situe *avant* la réalisation d'une expérience à résultat aléatoire, et l'on tâche de cerner et modéliser au mieux les éventualités qui se présentent à nous.

2.1 Espaces de Probabilité

Un ***espace de probabilité*** est constitué de trois ingrédients :

1. Un ensemble Ω, souvent appelé ***Univers***, dont les éléments ω représentent des résultats possibles de l'expérience.
2. Un ensemble $\mathcal{F}$ de sous-ensembles de Ω appelés ***événements*** et satisfaisant les axiomes **(A1)-(A2)-(A3)** donnés ci-dessous.
3. Une application $\mathbb{P} : \mathcal{F} \longrightarrow [0;1]$ affectant à chaque événement $E \in \mathcal{F}$ une mesure de probabilité $\mathbb{P}(E)$ située entre 0 et 1. Cette application $\mathbb{P} : \mathcal{F} \longrightarrow [0;1]$ est appelée ***mesure de probabilité*** et doit satisfaire les axiomes **(A4)-(A5)-(A6)** donnés ci-après.

On aura bien souvent affaire à un univers Ω qui est fini, et dans ce cas on pourra envisager un ***modèle équiprobable*** sur Ω :

- $\mathcal{F} = \mathcal{P}(\Omega)$ (tout sous-ens. de Ω constitue un événement).
- $\forall E \in \mathcal{P}(\Omega), \quad \mathbb{P}(E) = \frac{|E|}{|\Omega|}$ (la mesure d'un év. est *proportionnelle à sa cardinalité*).

Cependant, dans le cas général il se peut que l'espace Ω soit infini : il est alors impossible d'assigner à chaque événement une probabilité proportionnelle à sa cardinalité, et l'on peut aussi se trouver obligé de considérer une tribu d'événements $\mathcal{F}$ plus petite que $\mathcal{P}(\Omega)$ tout entier (en langage mathématique, on aura $\mathcal{F} \subsetneq \mathcal{P}(\Omega)$).
Quelles sont donc les *exigences minimales* requises pour n'importe quel espace de probabilité ?

Axiomes portant sur la Tribu d'événements $\mathcal{F}$:

(A1) $\emptyset$ et Ω doivent être des événements, appelés respectivement l'***événement impossible*** et l'***événement certain***. En langage mathématique, il faut que $\mathcal{F}$ vérifie $\emptyset \in \mathcal{F}$ et $\Omega \in \mathcal{F}$ (notation alternative : $\mathcal{F} \ni \emptyset$, $\mathcal{F} \ni \Omega$).

(A2) Il faut que la tribu $\mathcal{F}$ soit ***stable par passage au complémentaire*** : le complémentaire d'un événement E sera toujours un événement, appelé son événement contraire
(en langage mathématique : $\forall E \in \mathcal{F}$, $\overline{E} \in \mathcal{F}$).

(A3) Il faut que la tribu $\mathcal{F}$ soit ***stable pour les réunions ou intersections de suites d'événements*** : si (E_n) est une suite (finie ou infinie) d'événements la réunion $\bigcup_n E_n$ de tous les termes de la suite doit encore être un événement, et l'intersection $\bigcap_n E_n$ aussi.

Commençons tout de suite de voir un exemple fondamental d'espace de probabilité.
Exemple Fondamental :
Alice et Bob s'apprêtent à jouer à indéfiniment à "Pile ou Face" en utilisant une pièce de 1 Euro. Comment donner un modèle mathématique de *tous les déroulements possibles* pour cette expérience ?
On pourra représenter chaque 'Pile' obtenu par un 0, et chaque 'Face' par un 1 ; un résultat envisageable pour cette expérience est alors une suite infinie $\omega = (\omega_n)_{n\geq 1}$ à termes dans $\{0;1\}$. L'univers Ω est alors l'ensemble de toutes ces suites : $\Omega = \{0;1\}^{\mathbb{N}}$.
Pour des raisons profondes et mathématiquement subtiles, il s'avère impossible ici de proposer un modèle raisonnable en posant $\mathcal{F} = \mathcal{P}(\Omega)$. On pourra cependant envisager un espace de probabilité où *tous les sous-ensembles "raisonnables" de* Ω *figurent dans la tribu* $\mathcal{F}$.
Par exemple, *"Le 1er jet donne 'Pile'"* sera un événement E_1 ; mathématiquement :

$$E_1 = \{\omega \in \Omega | \, \omega_1 = 0\}, \quad E_1 \in \mathcal{F}$$

Plus généralement, on pourra considérer l'événement

$$E_k = \{\omega \in \Omega | \, \omega_k = 0\}$$

(en français : E_k *: "Le kème jet donne 'Pile'"*).
L'axiome **(A3)** nous permet alors d'affirmer que

$$E = \bigcup_{k\geq 1} E_k$$

figure aussi dans la tribu $\mathcal{F}$ (en français : E *: "'Pile' finit par apparaître"* est aussi un événement).
Quelle mesure de probabilité peut-on raisonnablement affecter à cet événement E ?

Voyons quels sont les axiomes **(A4)-(A5)-(A6)** que tout mesure de probabilité doit satisfaire :
Axiomes portant sur la Mesure de Probabilité $\mathbb{P}$:

(A4) ***Normalisation :*** on aura toujours $\mathbb{P}(\Omega) = 1$.

(A5) ***Additivité :*** si A et B sont deux événements *incompatibles*, autrement dit si $A \in \mathcal{F}, B \in \mathcal{F}$ et $A \cap B = \emptyset$, on aura toujours $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

(A6) ***Continuité :*** si $(F_k)_{k\geq 1}$ est une suite croissante d'événements, autrement dit si

$$F_1 \in \mathcal{F}, F_2 \in \mathcal{F}, \ldots, F_k \in \mathcal{F}, \ldots \text{ et } F_1 \subset F_2 \subset \ldots \subset F_k \subset F_{k+1} \subset \ldots$$

en posant $F = \lim_{k\to\infty} F_k = \bigcup_{k\geq 1} F_k$ on aura toujours

$$\mathbb{P}(F) = \mathbb{P}\left(\lim_{k\to\infty} F_k\right) = \lim_{k\to\infty} \mathbb{P}(F_k)$$

Les axiomes **(A4)-(A5)** sont tout à fait naturels, et dans le contexte d'un univers Ω fini ces deux seuls axiomes suffisent à définir ce qu'est une mesure de probabilité sur $(\Omega; \mathcal{F})$.
Cependant, dès lors que l'univers Ω est infini, l'axiome de continuité **(A6)** devient indispensable ! Voyons-en tout de suite une application concrète :

Exemple Fondamental (Reprise) :
Pour chaque entier $k \geq 1$, soit F_k : *"'Pile' apparaît au moins une fois lors des k premiers jets"*.
Mathématiquement, on a : $F_k = \cup_{i=1}^{k} E_i$; d'après l'axiome **(A3)**, F_k est donc un événement.
D'autre part, il est manifeste que $F_1 \subset F_2 \subset \ldots \subset F_k \subset F_{k+1} \subset \ldots$
D'après l'axiome de continuité **(A6)**, en posant $F = \lim_{k\to\infty} F_k = \bigcup_{k\geq 1} F_k$ on a donc

$$\mathbb{P}(F) = \lim_{k\to\infty} \mathbb{P}(F_k)$$

Remarquons qu'en français, l'événement F est simplement défini par :
"'Pile' finit par apparaître au cours du jeu".
Comment 'evaluer sa probabilité ? Il suffit pour cela de savoir évaluer les probabilités $\mathbb{P}(F_k)$, pour ensuite passer à la limite.
A supposer que la pièce utilisée *n'est pas biaisée*, on a pour chaque lancer :

$$\mathbb{P}\{\text{'Pile'}\} = \mathbb{P}\{\text{'Face'}\} = \frac{1}{2}$$

En remarquant que

$$\mathbb{P}(\overline{F_k}) = (\frac{1}{2})^k$$

et que

$$\mathbb{P}(F_k) + \mathbb{P}(\overline{F_k}) = 1$$

(d'après l'axiome **(A5)**), il vient

$$\mathbb{P}(F_k) = 1 - \mathbb{P}(\overline{F_k}) = 1 - \frac{1}{2^k}$$

Ainsi :

$$\mathbb{P}(F) = \lim_{k\to\infty} \mathbb{P}(F_k) = \lim_{k\to\infty} \left(1 - \frac{1}{2^k}\right) = 1$$

Plus généralement, si pour chaque lancer :

$$\mathbb{P}\{\text{'Pile'}\} = p = 1 - \mathbb{P}\{\text{'Face'}\},$$

on obtiendra

$$\forall k \geq 1, \quad \mathbb{P}(F_k) = 1 - \mathbb{P}(\overline{F_k}) = 1 - (1-p)^k$$

puis :

$$\mathbb{P}(F) = \lim_{k\to\infty} \mathbb{P}(F_k) = \lim_{k\to\infty} \left(1 - (1-p)^k\right) = 1$$

Rappelons que d'après les Lois de De Morgan, si (A_n) est une suite quelconque d'événements :

$$\overline{(\cup_{n\geq 1} A_n)} = \cap_{n\geq 1}\overline{A_n}, \quad \overline{(\cap_{n\geq 1} A_n)} = \cup_{n\geq 1}\overline{A_n}$$

(le *complémentaire d'une intersection* vaut la *réunion des complémentaires*, et le *complémentaire d'une réunion* vaut l'*intersection des complémentaires*).
Voyons à présent quelques-unes des principales propriétés de la tribu d'événements $\mathcal{F}$ et de la mesure de probabilité $\mathbb{P}$:

(P1) : Monotonie : pour deux événements A et B, on a toujours

$$A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$$

(P2) : Passage au Complémentaire : pour un événement quelconque A, on a toujours

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A),$$

en particulier : $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 0$.

(P3) : σ-Additivité : Si $(A_n)_{n\geq 1}$ est une suite infinie d'événements deux à deux incompatibles, on a

$$\mathbb{P}\left(\cup_{n\geq 1} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(A_n)$$

En particulier, si $(A_n)_{n\geq 1}$ constitue un ***système complet d'événements***, c'est à dire une suite d'événements deux à deux disjoints ($A_m \cap A_n = \emptyset$ si $m \neq n$) dont la réunion vaut Ω ($\bigcup_{n\geq 1} A_n = \Omega$), on aura

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\cup_{n\geq 1} A_n\right) = \mathbb{P}(\Omega) = 1$$

Plus généralement, si $(A_n)_{n\geq 1}$ constitue un ***système complet d'événements*** et si B est un autre événement, on aura toujours :

$$\mathbb{P}(B) = \mathbb{P}\left(B \cap \Omega\right) = \mathbb{P}\left(B \cap \bigcup_{n\geq 1} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B \cap A_n)$$

(P4) : Règles de Poincaré : pour deux événements quelconques E et F (compatibles ou non) on aura toujours

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$$

Pour trois événements quelconques E, F et G, on aura toujours

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E \cap F) - \mathbb{P}(E \cap G) - \mathbb{P}(F \cap G) + \mathbb{P}(E \cap F \cap G)$$

Remarquons que

1. La règle de σ-Additivité est souvent énoncée comme axiome et, combinée avec les autres axiomes, permet de retrouver la continuité de la mesure $\mathbb{P}$: une mesure de probabilité $\mathbb{P} : \mathcal{F} \longrightarrow \Omega$ est *"continue par le haut"* au sens où pour toute suite croissante d'événements $(E_n)_{n\geq 1}$,
$$\mathbb{P}(\cup_{n\geq 1} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n),$$
"continue par le bas" au sens où pour toute suite décroissante d'événements $(E_n)_{n\geq 1}$,
$$\mathbb{P}(\cap_{n\geq 1} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n).$$

2. La ***formule de Poincaré*** peut être énoncée en français dans toute sa généralité :
pour calculer la probabilité d'une réunion finie d'événements, on pourra former toutes les intersections possibles de ces événements, évaluer les probabilités correspondantes, puis affecter d'un signe $+$ les probabilités d'intersections impaires et d'un signe $-$ les probabilités d'intersections paires avant de sommer le tout !
Exemple avec quatre événements :
$$\begin{aligned}\mathbb{P}(E\cup F\cup G\cup H) \quad = \quad & \mathbb{P}(E)+\mathbb{P}(F)+\mathbb{P}(G)+\mathbb{P}(H)\\ & -\mathbb{P}(E\cap F)-\mathbb{P}(E\cap G)-\mathbb{P}(E\cap H)-\mathbb{P}(F\cap G)-\mathbb{P}(F\cap H)-\mathbb{P}(G\cap H)\\ & +\mathbb{P}(E\cap F\cap G)+\mathbb{P}(E\cap F\cap H)+\mathbb{P}(E\cap G\cap H)+\mathbb{P}(F\cap G\cap H)\\ & -\mathbb{P}(E\cap F\cap G\cap H)\end{aligned}$$

Exemples célèbres d'applications :

1. **Question du Chevalier de Méré à Blaise Pascal :**
Est-il plus probable de voir au moins un '6' apparaître lors de quatre jets d'un seul dé, ou de voir au moins un 'Double 6' apparaître lors de vingt-quatre jets de deux dés ?

2. **Le "Paradoxe des Anniversaires" :**
N étudiants viennent assister à leur premier cours dans le Grand Amphithéâtre de l'Université de Mandrika. Le professeur, quelque peu facétieux, demande de faire circuler une liste où chaque étudiant inscrit son anniversaire (jour du mois, sans l'année de naissance).

   a) Proposer un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ permettant de modéliser cette expérience aléatoire.

   b) On s'intéresse à l'événement

   E : "Il y a deux étudiants au moins dont les anniversaires sont identiques".

   Evaluer $\mathbb{P}(E)$ en fonction de l'entier N.
   Pour quelle valeur de N a-t-on $\mathbb{P}(E) \approx \frac{1}{2}$?

2.2 Rappels de Combinatoire

Comme évoqué plus haut, on sera souvent amené à utiliser un *modèle équiprobable*, où l'univers Ω est fini et où la probabilité d'un événement $E \subset \Omega$ est toujours proportionnelle à sa taille (à son nombre d'éléments) : $\mathbb{P}(E) = \frac{|E|}{|\Omega|}$.
Il y a donc des situations où l'évaluation de $\mathbb{P}(E)$ nécessite un certain "Savoir Compter" ou "Savoir-Faire Combinatoire". Avant de compter toutes les façons de réaliser un événement E, on pourra se poser les questions suivantes :

1. Quel est le nombre n d'objets de référence ?
2. Quel est le nombre p d'objets nécessaires à la réalisation de l'événement E ?
3. Ces p objets sont-ils considérés sans ordre ("en vrac", come lors d'un tirage simultané) ou avec ordre (les situations changent selon l'ordre d'apparition de p objets donnés) ?
4. Les répétitions sont-elles impossibles (p objets tous distincts, tirage sans remise) ou bien sont-elles possibles (comme lors d'un tirage avec remise) ?

Il conviendra donc de ne pas se lancer tête baissée dans des formules toutes faites.
Par exemple, on pourra s'appuyer sur un arbre de choix dans certaines situations où l'ordre d'apparition des objets compte, mais pas dans toutes les situations (l'utilisation d'un arbre de choix induit un ordre dans l'apparition des objets).
Par ailleurs, mentionnons que dans une situation comportant plusieurs choix ou plusieurs possibilités, on sera souvent amené à effectuer

- un *produit* lorsque la réalisation de l'événement requiert un choix, *suivi* d'un autre ...
- une *somme* lorsque cette réalisation comporte une possibilité, *ou bien* une autre ...

Voici maintenant une liste de coefficients combinatoires souvent rencontrés dans la pratique :

1. ***Tirages sans répétition et sans ordre :*** on utilisera souvent les ***coefficients binomiaux*** :
   si n et p sont deux entiers tels que $0 \leq p \leq n$, le coefficient binomial $C_n^p = \binom{n}{p}$ est donné par
   $$C_n^p = \binom{n}{p} = \frac{n!}{(n-p)!p!}$$
   En fait, $\binom{n}{p}$ compte le nombre de sous-ensembles à p éléments dans un ensemble comportant n éléments.
   Ces coefficients binomiaux satisfont les propriétés élémentaires suivantes :
   (i) $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{1} = \binom{n}{n-1} = n$ (*"un seul ensemble vide, un seul ensemble plein"*).
   (ii) $\binom{n}{p} = \binom{n}{n-p}$ (*"comptage des complémentaires"*)
   (iii) $\binom{n}{p} + \binom{n}{p+1} = \binom{n+1}{p+1}$ (*"Triangle de Pascal"*)

2. ***Tirages sans répétition et avec ordre :***
   on utilisera souvent les ***nombres d'arrangements*** A_n^p, définis pour $0 \leq p \leq n$ par

$$A_n^p = \frac{n!}{(n-p)!} = n(n-1)\dots(n-p+1) = p! \times C_n^p$$

   En fait, A_n^p compte le nombre de suites finies ayant p termes tous distincts pris dans un ensemble à n éléments.
   Dans le cas particulier où $p = n$, on obtient $A_n^n = n!$, on dit que l'on compte ainsi les ***permutations*** des n éléments de l'ensemble donné.

3. ***Tirages avec répétitions et avec ordre :***
   le comptage du nombre de suites à p termes pris dans un ensemble à n éléments et en admettant les répétitions se fait simplement en élevant n à la puissance p :

$$\sharp\{\text{Suites à } p \text{ termes pris dans } [\![1;n]\!]\} = n^p$$

   Notons qu'ici on peut tout à fait avoir $p > n$.

4. ***Tirages avec répétitions et sans ordre :***
   Une urne contient n boules numérotées de 1 à n, on effectue p tirages avec remise dans cette urne puis on compte combien de fois chacune des boules a été tirée. Le nombre de résultats possibles pour cette expérience est alors :

$$K_n^p = \binom{n+p-1}{p}$$

   Remarquons qu'ici encore, on peut tout à fait avoir $p > n$.

Le tableau qui suit offre une récapitulation des différents coefficients combinatoires rencontrés dans différentes situations de comptage :

Tirages	***Avec répétition***	***Sans répétition***
Avec ordre	n^p	A_n^p
Sans ordre	$\binom{n+p-1}{p}$	$\binom{n}{p}$

Voyons encore quelques exemples particulièrement courants où surgissent ces coefficients.

Exemples célèbres d'applications :

1. **Jeu de Loterie :**
   La compagnie de jeux "KIVATOUPERDRE" propose à ses clients des bulletins où l'on doit cocher cinq cases à choisir parmi les cases '1' à '99'. A la fin de la semaine, les cinq numéros gagnants sont tirés au sort en direct sur "BLINGTV".

a) Si les seuls bulletins gagnants sont ceux où sont cochés les cinq numéros gagnants, quels sont les chances de sucès d'un joueur ayant acheté un seul bulletin ?

b) Les règles du jeu sont assouplies, et un bulletin comportant quatre des cinq numéros gagnants donne aussi lieu à une certaine rémunération. Dans ces nouvelles conditions, quelles sont les chances de succès de notre joueur ?

2. **Problème des Chapeaux :**
10 comtesses anglaises sont invitées à une soirée mondaine dans un château ; chacune vient avec un très beau chapeau et confie son précieux couvre-chef au majordome avant de passer à table. Cependant, en plein milieu du dîner, un violent incendie se déclare, obligeant nos comtesses à fuir en se saisissant d'un chapeau quelconque en toute hâte.

a) Quelle est la probabilité que chaque comtesse se soit saisi de son propre chapeau ?

b) Quelle est la probabilité qu'une comtesse seulement ait réussi à se saisir de son propre chapeau ?

3. **Plaques d'Immatriculation :**
Les nouvelles plaques d'immatriculation françaises comportent deux lettres (caractères latins, en majuscules) suivies de trois chiffres puis de deux lettres. Combien de plaques distinctes peut-on élaborer suivant cette nouvelle norme ?

4. **Investissements d'une Entreprise :**
Une entreprise dispose d'un capital de n millions d'Euros, qu'elle décide d'investir sur quatre pôles d'activités A, B, C, D, par tranches de $100'000$ Euros. Combien de politiques d'investissement sont possibles :
- si l'entreprise décide d'investir la totalité des n millions d'Euros en réserve ?
- si l'entreprise décide de conserver éventuellement une partie des n millions d'Euros en réserve ?

2.3 Probabilités Conditionnelles, Evénements Indépendants, Formule de Bayes

Considérons un espace de probabilité $(\Omega; \mathcal{F}; \mathbb{P})$, puis deux événements $A, B \subset \Omega$ tels que $\mathbb{P}(B) > 0$. On définit la ***probabilité conditionnelle de*** A ***sachant*** B par

$$\mathbb{P}_B(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

En fixant l'"événement conditionnant" B et en faisant varier A dans la tribu $\mathcal{F}$, on obtient ainsi une nouvelle mesure de probabilité $\mathbb{P}_B : \mathcal{F} \longrightarrow [0; 1]$. Cette nouvelle mesure est telle que

$$\begin{cases} \mathbb{P}_B(A) = 0 & \text{si } A \cap B = \emptyset \\ \mathbb{P}_B(A) = 1 & \text{si } B \subset A \end{cases}$$

Toutes les règles valables pour une mesure de probabilité quelconque sur $(\Omega; \mathcal{F})$ s'appliquent donc à cette mesure $\mathbb{P}_B$. En outre :

1. **Probabilité d'une intersection :**
   si les événements $B_1, B_2, \ldots, B_k$ vérifient $\mathbb{P}(B_1 \cap B_2 \cap \ldots \cap B_k) > 0$, on pourra écrire
   $$\mathbb{P}(B_1 \cap B_2 \cap \ldots \cap B_k) = \mathbb{P}(B_1) \cdot \mathbb{P}_{B_1}(B_2) \cdot \mathbb{P}_{B_1 \cap B_2}(B_3) \cdot \ldots \cdot \ldots \cdot \mathbb{P}_{B_1 \cap B_2 \cap \ldots \cap B_{k-1}}(B_k)$$

2. **Utilisation d'un système complet d'événements :**
   Si les événements $B_1, B_2, \ldots, B_N$ constituent un système complet d'événements non-négligeables, on aura pour un événement E quelconque :
   $$\begin{array}{rcl} \mathbb{P}(E) & = & \sum_{n=1}^{N} \mathbb{P}(E \cap B_n) \\ & = & \mathbb{P}(E \cap B_1) + \mathbb{P}(E \cap B_2) + \ldots + \mathbb{P}(E \cap B_N) \\ & = & \mathbb{P}_{B_1}(E)\mathbb{P}(B_1) + \mathbb{P}_{B_2}(E)\mathbb{P}(B_2) + \ldots + \mathbb{P}_{B_N}(E)\mathbb{P}(B_N) \\ & = & \sum_{n=1}^{N} \mathbb{P}_{B_n}(E)\mathbb{P}(B_n) \end{array}$$
   Pour un système complet à deux événements seulement B et $\overline{B}$, en supposant que B et $\overline{B}$ sont tout deux non-négligeables, on a donc
   $$\mathbb{P}(E) = \mathbb{P}_B(E)\mathbb{P}(B) + \mathbb{P}_{\overline{B}}(E)\mathbb{P}(\overline{B})$$

3. **Formule de Bayes :**
   dans le contexte du point précédent, en supposant en outre que E est lui aussi non-négligeable, on a
   $$\mathbb{P}_E(B_k) = \frac{\mathbb{P}_{B_k}(E)\mathbb{P}(B_k)}{\sum_{n=1}^{N} \mathbb{P}_{B_n}(E)\mathbb{P}(B_n)}$$
   pour tout indice $k \in [\![1; N]\!]$.
   Dans le cas particulier d'un système complet à deux événements non-négligeables B et $\overline{B}$, on a donc
   $$\mathbb{P}_E(B) = \frac{\mathbb{P}_B(E)\mathbb{P}(B)}{\mathbb{P}_B(E)\mathbb{P}(B) + \mathbb{P}_{\overline{B}}(E)\mathbb{P}(\overline{B})}$$

Cette dernière formule, découverte au XVIIIème siècle par le Révérend Thomas Bayes, s'est avérée particulièrement utile puisqu'elle permet de ***renverser des conditionnements*** !
En voici deux applications célèbres :

Exemples célèbres d'applications de la Formule de Bayes :

1. **Test Médicaux :**
   Un test médical est mis au point dans le but de détecter une maladie d'origine virale. On sait que 1% de la population considérée est atteint de cette maladie, tandis que
   $$\alpha = \mathbb{P}(\text{“Test positif“}|\text{”Patient Malade”}) = 0,95 = 1 - \mathbb{P}(\text{“Test négatif“}|\text{”Patient Malade”})$$
   et
   $$\beta = \mathbb{P}(\text{“Test négatif“}|\text{”Patient Sain”}) = 0,94 = 1 - \mathbb{P}(\text{“Test positif“}|\text{”Patient Sain”})$$
   (On dit que la ***sensibilité*** α du Test vaut 95%, tandis que sa ***spécificité*** β vaut 96%).
   Un Patient choisi au hasard dans la population est testé positivement.
   Quelle est la probabilité qu'il soit effectivement malade??

2. **Jeu de "Monty Hall" :**
   Un sémillant animateur de télévision reçoit un joueur sur son plateau. Sur ce plateau figurent trois portes A, B, C, et l'une des trois portes cache une authentique Cadillac, tandis que les deux autres ne cachent rien du tout.
   L'animateur sait évidemment où se trouve la Cadillac, il invite le joueur à désigner l'une des trois portes et lui promet qu'ensuite, il ouvrira l'une des deux autres portes pour lui montrer qu'il n'y a rien derièrre.
   Le joueur désigne la porte A, et notre présentateur lui ouvre ensuite la porte B avec un grand sourire pour lui montrer qu'elle ne cachait rien, lui laissant la décision de maintenir son premier choix (porte A) ou de reporter son choix sur la porte restante (porte C).
   Vérifier qu'un changement de porte donne ***deux fois plus de chances*** au joueur de remporter la mise qu'une conservation de la porte initiale.

Voici enfin une dernière définition particulièrement importante :
deux événements non-négligeables A et B seront dits ***indépendants*** lorsque

$$\mathbb{P}_B(A) = \mathbb{P}(A) \text{ et } \mathbb{P}_A(B) = \mathbb{P}(B)$$

Alternativement, on dira que les événements A et B sont ***indépendants*** si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

(Avantage de cette définition alternative : pas de restriction du type "A et B non-négligeables").
Plus généralement, trois événements A, B, C seront dits indépendants si

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B),\ \mathbb{P}(A\cap C) = \mathbb{P}(A)\cdot\mathbb{P}(C),\ \mathbb{P}(B\cap C) = \mathbb{P}(B)\cdot\mathbb{P}(C),\ \mathbb{P}(A\cap B\cap C) = \mathbb{P}(A)\cdot\mathbb{P}(B)\cdot\mathbb{P}(C)$$

Lorsque les trois premières identités ci-dessus sont vraies, on parlera d'***événements indépendants deux à deux***. Pour conclure, remarquons que

1. Trois événements peuvent être indépendants deux à deux sans être indépendants dans leur ensemble.
   Par exemple, si l'on jette trois dés simultanément (de couleurs bleue, rouge et verte) et si les événements A, B et C sont définis par
   A : "Les dés bleu et rouge donnent le même chiffre",
   B : "Les dés bleu et vert donnent le même chiffre",
   C : "Les dés rouge et vert donnent le même chiffre",
   on vérifie sans trop de peine que A, B, C sont deux à deux indépendants sans être indépendants dans leur ensemble.
   Insistons cependant sur le fait que l'expression "E, F, G sont indépendants" signifie qu'ils le sont au sens le plus fort (dans leur ensemble).

2. si A ou $\overline{A}$ est négligeable, A est indépendant de tout autre événement B dans Ω.

3. Il faut bien se garder de confondre les notions d'***indépendance*** et d'***incompatibilité*** : si deux événements A et B sont incompatibles (ou disjoints, i.e. tels que $A \cap B = \emptyset$), on aura $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$, et donc $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$, à moins que l'un de ces deux événements soit négligeable.

2.4 Exercices d'Application

I Un peu d'Analyse Combinatoire

1. Combien peut-on former de sigles d'entreprises ayant :
   - exactement quatre lettres de l'alphabet latin ?
   - au plus quatre lettres de l'alphabet latin ?
2. La belote se joue avec 32 cartes, chaque joueur recevant initialement une main de cinq cartes. Combien de mains différentes peut obtenir un joueur
   - au total ?
   - avec exactement un as ?
   - avec au moins un as ?
   - avec exactement un as et un roi ?
   - avec deux cartes d'une couleur et trois d'une autre ?
3. Une équipe de volley-ball de six personnes reçoit au hasard 6 maillots numérotés de 1 à 6. Quel est le nombre de répartitions possibles des maillots ?
   Si l'un des joueurs tient absolument à porter le No. '1', combien reste-t-il de possibilités ?
4. On considère douze patients : six d'entre eux sont atteints d'une maladie M, et les six autres ne le sont pas. On répartit ces patients en deux groupes A et B de six patients chacun.
   Quelle est la probabilité que chaque groupe ne contienne qu'un seul type d'individu (que des malades ou que des gens sains) ?
   Quelle est la probabilité que chaque groupe comporte trois individus de chaque type (trois patients malades et trois patients sains) ?
5. Le poker se joue avec 32 cartes, chaque joueur recevant initialement une main de cinq cartes. Un *"Brelan"* est une main comportant trois cartes de même valeur suivies de deux cartes de valeurs distinctes, tandis qu'un *"Full"* est une main comportant trois cartes de même valeur suivies d'une paire; enfin, un *"Carré"* est une main comportant quatre cartes de même valeur suivies d'une autre carte. Quel est le nombre de façons possibles d'obtenir
   - un Carré ?
   - un Full ?
   - un Brelan ?

   En déduire la probabilité pour un joueur de recevoir initialement un Carré, un Full ou un Brelan.
6. On considère k urnes contenant chacune n boules identiques numérotées de 1 à n. Un joueur extrait au hasard une boule de chaque urne. Pour un entier fixé m tel que $1 \leq m \leq n$, quelle est la probabilité que le plus grand nombre ainsi obtenu soit m ?

II Espaces de Probabilité

Pour chacune des questions de cet exercice, on cherchera non seulement à évaluer correctement les probabilités demandées, mais aussi à trouver un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ correspondant à l'expérience considérée.

1. On jette trois fois un dé équilibré; calculer la probabilité d'obtenir :
   - au moins un '6',
   - un '6' exactement,
   - au moins deux résultats identiques,
   - une somme de résultats paire.
2. Alice et Bob décident de jouer de la façon suivante : Alice lance deux dés équilibrés à six faces, tandis que Bob lance un dé équilibré à douze faces, et celui qui obtient le plus grand résultat emporte la mise (match nul si les deux résultats sont égaux). Ce jeu est-il équilibré ? On calculera la probabilité qu'Alice gagne, ainsi que celle d'un match nul.
3. Un savant fou permute au hasard les 100 livres de sa bibliothèque; quelle est la probabilité que les livres 1 et 2 se retrouvent côte à côte dans le bon ordre?
4. n marins ivres regagnent leur bateau de nuit.
   - Que vaut la probabilité p_n qu'aucun d'eux ne dorme dans son propre hamac ?
   - Trouver la limite de p_n pour $n \to +\infty$.

III Encadrements et Evaluations de Probabilités

Dans l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, on considère deux événements A et B tels que $\mathbb{P}(A) = \frac{1}{2}$ et $\mathbb{P}(B) = \frac{1}{4}$.

1. Donner un encadrement de $\mathbb{P}(A \cup B)$ et de $\mathbb{P}(A \cap B)$.
2. Déterminer la valeur de $\mathbb{P}(A \cup B)$ lorsque A et B sont disjoints, puis lorsque $\mathbb{P}(A \cap B) = \frac{1}{5}$, et enfin lorsque A et B sont indépendants.

On se donne à présent deux autres événements C et D, et l'on suppose que leurs probabilités respectives sont les deux solutions de l'équation

$$10x^2 - 9x + 2 = 0,$$

avec en outre $\mathbb{P}(C) < \mathbb{P}(D)$.

3. Déterminer $\mathbb{P}(C)$ et $\mathbb{P}(D)$.
4. Calculer $\mathbb{P}(C \cup D)$ et de $\mathbb{P}(C \cap D)$ lorsque C et D sont disjoints.
5. Calculer $\mathbb{P}(C \cup D)$ et de $\mathbb{P}(C \cap D)$ lorsque C et D sont indépendants.

IV Encadrements et Evaluations de Probabilités (suite)

On suppose que les événements $A, B, C \subset \Omega$ sont tels que

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{3}, \quad \mathbb{P}(C) = \frac{1}{4}.$$

1. Ces événements peuvent-ils être disjoints deux à deux ?
2. On suppose que A et B sont indépendants et que A et C le sont aussi, ainsi que B et C. Evaluer $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cup C)$ et $\mathbb{P}(B \cup C)$.
   Peut-on, dans ces conditions, évaluer $\mathbb{P}(A \cup B \cup C)$?

V Définition des Probabilités Conditionnelles

1.*a*) Mon voisin a deux enfants dont une fille. Quelle est la probabilité que l'autre enfant soit un garçon ?

1.*b*) Mon voisin a deux enfants, et le plus jeune des deux est une fille. Quelle est la probabilité que l'aîné soit un garçon ?

2. On considère trois cartes, l'une ayant deux faces rouges, la deuxième deux faces blanches et la troisième une face rouge et une face blanche.
   Un joueur tire une carte au hasard, puis il expose l'une des deux faces de la carte tirée au hasard; cette face est rouge. Parieriez-vous que la face cachée est blanche ? Pour prendre la bonne décision, on pourra
   a) Donner un espace de probabilité correspondant à cette expérience.
   b) Calculer la prob. que la face cachée soit blanche sachant que la face exposée est rouge.
3. La soeur et la femme de Bob ont les yeux bleus, mais ses parents ont les yeux marrons.
   a) Quelle est la probabilité que Bob ait les yeux bleus ?
   b) Quelle est la probabilité que le 1er enfant de Bob ait les yeux bleus sachant que Bob lui-même a les yeux marrons ?
   c) Quelle est la probabilité que le 2ème enfant de Bob ait les yeux bleus sachant que son 1er enfant a les yeux marrons ? Commentaire ?
4. Dans une résidence de vacances, deux activités sportives sont proposées : le rafting et la randonnée. Sur les 120 résidents présents, 80 ont fait de la randonnée, 50 ont fait du rafting et 20 n'ont pratiqué aucune des deux activités sportives. On choisit un résident au hasard.
   a) Quelle est la probabilité qu'elle ait pratiqué un sport ?
   b) Sachant que la personne a fait du rafting, quelle est la probabilité qu'elle ait aussi fait de la randonnée ?
   c) Sachant que la personne a fait du sport, quelle est la probabilité qu'elle ait participé à la randonnée ?
5. Une enquête est menée auprès de 1000 usagers du réseau de la RATP, qui se voient tous poser la question suivante :
   "Au cours des deux derniers mois, combien de fois êtes-vous arrivé en retard au travail ?"
   Les réponses ont été regroupées dans le tableau ci-dessous :

Retards 1er mois / Retards 2ème mois	*0*	*1*	*2 ou plus*	*Total*
0	*262*	*212*	*73*	*547*
1	*250*	*73*	*23*	*346*
2 ou plus	*60*	*33*	*14*	*107*
Total	*572*	*318*	*110*	*1000*

(a) Un individu de cette population est choisi au hasard. Quelle est la probabilité qu'il ait eu au moins un retard durant le premier mois ?

(b) Dans les mêmes conditions, quelle est la probabilité qu'un individu ait eu au moins un retard durant le deuxième mois sachant qu'il n'en a pas eu durant le premier mois ?

VI Formule de Bayes

1. Un enfant joue avec cent dés, dont 25 sont pipés; pour ces derniers, la probabilité d'obtenir un $'6'$ vaut 1/3 au lieu de 1/6. L'enfant choisit un dé au hasard.
   a) Calculer la probabilité qu'en le lançant, il obtienne un $'6'$.
   b) En déduire la probabilité que le dé soit pipé sachant qu'en le lançant, il a obtenu un $'6'$.

2. Une enquête économétrique est conduite dans la région $PACA$, qui comporte 70% de PME et 30% de Grandes Entreprises; on remarque alors que 5% des PME de la région font faillite chaque année, tandis que ce pourcentage n'est que de 1% pour les GE.
   Une entreprise de Marseille fait faillite; quelle est la probabilité que cette entreprise soit une PME ?

3. Lors d'un examen oral de Mathématiques, un candidat est amené à choisir une question au hasard dans un lot de cent questions, 40 d'entre elles étant des questions de géométrie et 60 d'entre elles des questions d'analyse. On estime qu'il a 3 chances sur 4 de répondre correctement à une question de géométrie, et 2 chances sur 3 de répondre correctement à une question d'analyse.
   a) Quelle est la probabilité qu'a le candidat de répondre correctement à la question qu'il a choisie ?
   b) Sachant que le candidat répond correctement à la question choisie, quelle est la probabilité qu'il ait eu affaire à une question d'analyse ?

4. Chaque année, les équipes d'aviron de Cambridge et d'Oxford s'affrontent lors d'une course au début de l'été. On considère qu'en Angleterre, en cette saison, il fait très mauvais avec probabilité $0,2$, mauvais avec probabilité $0,5$ et beau avec probabilité $0,3$. Par très mauvais temps, Cambridge est susceptible de gagner avec probabilité $0,8$, par mauvais temps cette probabilité n'est plus que de $0,6$ et par beau temps elle descend à $0,2$ (la course ne se solde jamais par une égalité des deux équipes).
   M. Smith, en voyage à Tombouctou, apprend qu'Oxford vient de remporter cette

course d'aviron. Avec quelle probabilité a-t-il fait beau en Angleterre le jour de la course ?

VII Cogitations du Joueur de Bridge

Le Bridge se joue à quatre joueurs avec un jeu de 52 cartes, et chaque joueur reçoit une main de 13 cartes.

Le malheureux Bob reçoit une main ne comportant pas d'As; calculer la probabilité que sa partenaire Alice reçoive une main

1. Sans As. 2. Comportant au moins deux As.

VIII Loterie et Probabilités composées

Une urne contient n boules identiques numérotées de 1 à n. On tire, sans relise, p boules de cette urne (où $1 \leq p \leq n$).

Pour un entier k fixé entre 1 et n, quelle est la probabilité que tous les numéros tirés soient $\leq k$?

IX Oenologie et Fiabilité

n oenologues réputés se voient servir le même vin lors d'une dégustation "à l'aveugle", et savent seulement que celui-ci est un Bordeaux avec prob. $\frac{1}{2}$ ou un vin de Californie avec prob. $\frac{1}{2}$. Chacun d'eux reconnaît le vin sans se tromper avec probabilité $p = 0,75$, et ces oenologues rendent leurs jugements de manière indépendante.

Sachant que les $(n-1)$ premiers dégustateurs ont cru reconnaître un Bordeaux et que seul le dernier a cru reconnaître un vin de Californie, quelle est la probabilité que le vin servi soit effectivement californien ?

Et si $p = 0,49$?

X Questionnaires à Choix Multiples

1. On considère l'intervalle discret $\Omega = [\![1; 10]\!] = \{1, 2, 3, \ldots, 10\}$ muni de sa mesure de probabilité uniforme (où chaque événement élémentaire a pour probabilité $0,1$) puis les événements $A, B, C \subset \Omega$ donnés par :

$$A = \{2, 4, 6, 8, 10\}, \quad B = \{3, 4, 5, 10\} \text{ et } C = \{1, 2, 3, 4, 5\}$$

QCM1 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (ou sont) exacte(s) ?

A $\mathbb{P}(A \cap B) = 0,2$
B $\mathbb{P}(A \cap B) = 0,3$
C $\mathbb{P}(A \cap C) = 0,2$
D $\mathbb{P}(A \cap C) = 0,3$
E $\mathbb{P}(B \cap \overline{C}) = 0,2$

QCM2 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (ou sont) exacte(s) ?

A $\mathbb{P}(A \cup B) = 0,7$
B $\mathbb{P}(A \cup B) = 0,8$
C $\mathbb{P}(A \cup C) = 0,7$
D $\mathbb{P}(A \cup C) = 0,8$
E $\mathbb{P}(B \cup \overline{C}) = 0,7$

QCM3 : Parmi les propositions suivantes, laquelle (ou lesquelles) est (ou sont) vraie(s) ?

A A et B sont indépendants.
B B et C sont indépendants.
C $(A \cap B)$ et $(B \cap C)$ sont indépendants.
D $\overline{(A \cap B)}$ et $(B \cap C)$ sont indépendants.
E $(A \cap B)$ et $\overline{C}$ sont indépendants.

QCM4 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (ou sont) exacte(s) ?

A $\mathbb{P}(A|B) = 0,5$
B $\mathbb{P}[(A \cap B)|C] = 0,2$
C $\mathbb{P}[(A \cap B)|(A \cup B)] = 0,4$
D $\mathbb{P}[A|(B \cup C)] = 0,4$
E $\mathbb{P}[(A \cap B)|(B \cup C)] = 0,25$

2. Dans un certain univers Ω, on considère deux événements A et B pour lesquels $\mathbb{P}(A) = 0,7$, $\mathbb{P}(A \cap B) = 0,4$ et $\mathbb{P}(A|B) = 0,8$.
QCM5 : Quelle est la valeur de $\mathbb{P}(B)$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$
E $0,5$

QCM6 : Quelle est la valeur de $\mathbb{P}(A|\overline{B})$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$
E $0,5$

QCM7 : Quelle est la valeur de $\mathbb{P}(\overline{A} \cap \overline{B})$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$
E $0,5$

3. On considère à présent deux événements C et D tels que $\mathbb{P}(C \cap D) = 0,1$, $\mathbb{P}(C \cup \overline{D}) = 0,6$ et $\mathbb{P}(\overline{C}|\overline{D}) = 0,8$.

QCM8 : Quelle est la valeur de $\mathbb{P}(D)$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$
E $0,5$

QCM9 : Quelle est la valeur de $\mathbb{P}(\overline{C} \cap \overline{D})$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$
E $0,5$

QCM10 : Quelle est la valeur de $\mathbb{P}(C)$?

A $0,1$
B $0,2$
C $0,3$
D $0,4$

E $0,5$

4. Dans un pays lointain, un individu naissant dans les années 1950 avait une probabilité égale à $0,95$ d'atteindre l'âge de 18 ans. S'il atteignait cet âge, il avait une probabilité égale à $0,1$ de se présenter au baccalauréat ; parmi les candidats au baccalauréat, il y avait 80% de reçus, puis parmi les titulaires du baccalauréat, 10% de jeunes choisissant de s'inscrire à la faculté de médecine. Enfin, un étudiant s'inscrivant en première année de médecine avait 10% de chances de devenir médecin.
QCM11 : Quelle était dans ces conditions la proportion de médecins parmi les individus nés dans les années 1950 dans ce pays ?

   A $0,000076$
   B $0,00076$
   C $0,0076$
   D $0,076$
   E $0,2$

5. On classe les douleurs aigües de la fosse iliaque droite en appendicite, colique néphrétique et autres causes. Ces trois causes sont supposées exclusives et exhaustives. Les prévalences de ces différentes causes dans un service de consultations d'urgence sont de $0,35$ pour l'appendicite, de $0,2$ pour la colique néphrétique et $0,45$ pour les autres causes. Des signes digestifs sont présents dans 50% des appendicites, mais aussi pour 30% des coliques néphrétiques et 40% des douleurs ayant d'autres causes.
QCM12 : Au total, quelle est la proportion de malades ayant des troubles digestifs parmi ceux qui se présentent à la consultation pour une douleur aigüe de la fosse iliaque droite ?

   A $0,333$
   B $0,405$
   C $0,415$
   D $0,551$
   E $0,768$

QCM13 : Si un malade se présente à la consultation avec une douleur aigüe de la fosse iliaque droite sans aucun trouble digestif, la probabilité pour qu'il soit atteint d'une appendicite vaut :

   A $0,299$
   B $0,333$
   C $0,422$
   D $0,544$
   E $0,611$

6. Un singe tape des lettres au hasard sur le clavier d'une machine à écrire.
QCM14 : Que vaut la probabilité qu'il tape le mot "oui" s'il ne tape que trois lettres à la suite ?

A $0,000043$
B $0,000057$
C $0,00043$
D $0,00057$
E $0,00067$

QCM15 : Le singe tape une série de $30'000$ lettres, que l'on découpe en $10'000$ mots de trois lettres. Quelle est la probabilité de voir figurer au moins une fois le mot "oui" parmi ces $10'000$ mots ?

A $0,043$
B $0,057$
C $0,43$
D $0,57$
E $0,67$

7. Une maladie M se présente sous deux formes cliniques, une forme modérée et une forme sévère. La proportion de malades atteints de la forme modérée vaut $0,7$. Les proportions de malades fébriles s'élèvent respectivement à $0,4$ parmi les malades atteints de la forme modérée et à $0,8$ parmi les malades atteints de la forme sévère.
QCM16 : Quelle est la proportion de malades fébriles parmi les individus atteints de la maladie M ?

A $0,12$
B $0,52$
C $0,56$
D $0,60$
E $0,68$

QCM17 : Si un malade atteint de la maladie M est fébrile, quelle est la probabilité qu'il soit atteint de la forme sévère ?

A $0,33$
B $0,46$
C $0,50$
D $0,52$
E $0,60$

8. On considère qu'une maladie M est causée principalement par deux facteurs connus, les facteurs A et B. Ces facteurs surviennent indépendamment chez les individus de la population étudiée, avec des probabilités respectives sélevant à $0,2$ pour le facteur A et $0,3$ pour le facteur B. Pour un individu affecté du facteur A mais pas du facteur B, la probabilité d'être atteint de la maladie M vaut $0,5$; cette probabilité s'élève à $0,6$ pour un individu affecté du facteur B mais pas du facteur A, à $0,8$ pour un individu affecté des deux facteurs et à $0,1$ chez un individu affecté d'aucun de ces

deux facteurs.

QCM18 : Quelle est la probabilité qu'un individu choisi au hasard soit atteint de la maladie M ?

A $0,308$
B $0,318$
C $0,342$
D $0,407$
E $0,558$

QCM19 : Si un individu choisi au hasard est atteint de la maladie M, quelle est la probabilité qu'il soit affecté du facteur A ?

A $0,371$
B $0,383$
C $0,407$
D $0,522$
E $0,558$

QCM20 : Si un individu est atteint de la maladie M et affecté du facteur A, quelle est la probabilité qu'il soit aussi affecté du facteur B ?

A $0,300$
B $0,342$
C $0,371$
D $0,383$
E $0,407$

3 VARIABLES ALEATOIRES DISCRETES

3.1 Définitions élémentaires, variables classiques

De manière informelle, on peut dire qu'une variable aléatoire X est une variable numérique qui est susceptible de prendre différentes valeurs suivant certaines probabilités.
Plus techniquement, une variable aléatoire est une application $X : \Omega \longrightarrow \mathbb{R}$ définie sur un espace de probabilité $(\Omega; \mathcal{F}; \mathbb{P})$ et à valeurs réelles. En outre, on demande à X de respecter la structure de l'espace $(\Omega; \mathcal{F}; \mathbb{P})$ au sens où pour tout intervalle $I \subset \mathbb{R}$:

$$X^{-1}(I) = \{\omega \in \Omega | \, X(\omega) \in I\}$$

doit être un événement dans Ω ($X^{-1}(I) \in \mathcal{F}$). L'ensemble des valeurs envisageables pour la variable aléatoire X peut donc être noté $X(\Omega)$.
On dira que X est une ***variable aléatoire discrète*** lorsque ces valeurs envisageables peuvent être rangées comme termes d'une suite, finie ou infinie :

$$X(\Omega) = \{v_1, v_2, \ldots, v_N\} \text{ ou } X(\Omega) = \{v_1, v_2, \ldots, v_n, v_{n+1}, \ldots\}$$

(En règle générale, ces valeurs possibles d'une v.a. discrète seront elles-mêmes des entiers).
Enfin, donner la ***loi*** de la variable aléatoire discrète X, c'est spécifier les probabilités

$$\mathbb{P}\{X = v_1\}, \mathbb{P}\{X = v_2\}, \ldots, \mathbb{P}\{X = v_k\}, \ldots$$

de voir X prendre telle ou telle autre valeur envisageable.
Voyons maintenant quelques exemples archi-classiques de variables aléatoires discrètes, exemples que l'on retrouvera souvent dans la pratique.
Il est important de noter que chacune de ces lois classiques est attachée à un certain type d'***expérience*** : en général, dans le contexte des variables discrètes, on aura affaire à des ***variables de comptage***, ou encore à des ***variables de temps d'attente***.

3.1.1 Variables de Bernoulli

Alice et Bob jouent à Pile ou Face. La pièce utilisée a ceci de particulier qu'elle produit 'Pile' avec probabilité p à chaque lancer, tandis que 'Face' sort avec probabilité $(1-p)$ à chaque lancer. Cette pièce est lancée une fois, et le résultat 'Pile' permet à Alice de gagner 1 Euro tandis que Bob ne gagne rien du tout ; dans le cas contraire (résultat 'Face'), c'est Alice qui ne gagne rien du tout tandis que Bob gagne 1 Euro.
Soit X la variable aléatoire représentant le gain d'Alice à l'issue de ce lancer. On a manifestement :

$$X(\Omega) = \{0; 1\}, \, p_0 = \mathbb{P}\{X = 0\} = (1-p), \, p_1 = \mathbb{P}\{X = 1\} = p$$

Une telle variable X est appelée ***Variable de Bernoulli de paramètre*** p.

3.1.2 Variables Binomiales

Imaginons à présent que cette même pièce est lancée n fois d'affilée et qu'Alice gagne 1 Euro pour chaque apparition de 'Pile', rien du tout pour chaque apparition de 'Face'. Y désignant les gains d'Alice à l'issue de ces n lancers, on a maintenant

$$Y(\Omega) = \{0; 1; 2; \ldots; n\} = [\![0; n]\!], \; p_0 = \mathbb{P}\{Y = 0\} = (1-p)^n, \; p_1 = \mathbb{P}\{Y = 1\} = np(1-p)^{n-1}, \ldots$$

Plus précisément, Y est une ***Variable Binomiale de paramètres*** n ***et*** p, sa loi est donnée par

$$\forall k \in [\![0; n]\!], \; p_k = \mathbb{P}\{Y = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

Par exemple, pour $n = 4$ et $p = \frac{1}{3}$ on obtient

$$\mathbb{P}\{Y = 0\} = (\frac{2}{3})^4 \approx 0,198, \; \mathbb{P}\{Y = 1\} = 4\frac{1}{3}(\frac{2}{3})^3 \approx 0,395, \; \mathbb{P}\{Y = 2\} = \binom{4}{2}(\frac{1}{3})^2(\frac{2}{3})^2 = 6(\frac{1}{3})^2(\frac{2}{3})^2 \approx 0,296$$

Notons tout de suite qu'une variable binomiale de paramètres n et p pourra toujours être considérée comme somme de n variables de Bernoulli *indépendantes* et de *même paramètre* p :

$$Y = X_1 + X_2 + \ldots + X_n$$

Dans notre exemple, X_1 représente le résultat du 1er jet, X_2 celui du 2ème jet, etc ....
Ces variables de Bernoulli sont ***indépendantes***, au sens où tous les événements du type

$$\{X_1 = v_1\}, \{X_2 = v_2\}, \{X_3 = v_3\}, \ldots, \{X_n = v_n\}$$

sont indépendants (validité des "Règles du Produit" avec la probabilité $\mathbb{P}$).
Intuitivement, cela correspond au fait que "les résultats des différents jets ne s'influencent pas mutuellement".
Une variable binomiale Y de paramètres n et p est donc une ***variable de comptage***, que l'on pourra tout aussi bien se représenter en termes de ***tirages au sort avec remise*** : une urne contient une proportion p de boules noires et une proportion $(1-p)$ de boules blanches, un joueur effectue n tirages au sort *avec remise* dans cette urne et compte le nombre Y de boules noires ainsi obtenues.

3.1.3 Variables Hypergéométriques

Ces variables aléatoires discrètes sont encore des ***variables de comptage***, que l'on pourra se représenter en termes de ***tirages au sort sans remise*** : une urne contient une proportion p de boules noires et une proportion $(1-p)$ de boules blanches, un joueur effectue n tirages au sort *sans remise* dans cette urne et compte le nombre Z de boules noires ainsi obtenues. Pour spécifier la loi de cette variable hypergéométrique Z, il nous faut aussi connaître le nombre N de boules contenues initialement dans l'urne ; il y a alors initialement pN boules noires et $(1-p)N$

boules blanches, et pour plus de commodité on supposera que le nombre n de tirages est majoré par ces nombres initiaux de boules noires et de boules blanches :

$$n \leq pN, \quad n \leq (1-p)N$$

La variable hypergéométrique Z est alors à valeurs dans l'intervalle discret $[\![0; n]\!]$ et telle que

$$\forall k \in [\![0; n]\!], \mathbb{P}\{Z = k\} = \frac{\binom{pN}{k}\binom{(1-p)N}{(n-k)}}{\binom{N}{n}}$$

Cette variable hypergéométrique Z de paramètres N, n et p n'a donc pas tout à fait la même loi que la variable binomiale Y de paramètres n et p.
Cependant les ensembles de valeurs envisageables pour Y et Z sont identiques ($Y(\Omega) = Z(\Omega) = [\![0; n]\!]$), et pour $N \longrightarrow +\infty$ à n et p fixés, cette différence de lois s'estompe :

$$\forall k \in [\![0; n]\!], \lim_{N\to\infty} \frac{\binom{pN}{k}\binom{(1-p)N}{(n-k)}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}$$

Ce résultat ardu en apparence devient facile à comprendre si l'on raisonne en termes de tirages au sort : dans une vaste urne contenant $N = 10'000$ bulletins dont $3'000$ bulletins noirs, les comptages du nombre de bulletins noirs obtenus lors d'un tirage de $n = 10$ bulletins ***avec*** ou ***sans remise*** ont des comportements quasiment identiques ! On dit alors que les lois Hypergéométrique de paramètres $N = 10'000$, $n = 10$, $p = 0,3$ et Binomiale de paramètres $n = 10$, $p = 0,3$ sont quasiment identiques :

$$\mathcal{H}(N = 10'000; n = 10; p = 0,3) \approx \mathcal{B}(n = 10; p = 0,3)$$

Plus généralement, on peut aussi écrire que

$$\mathcal{H}(N; n; p) \underset{N\to\infty}{\longrightarrow} \mathcal{B}(n; p)$$

ce qui constitue un premier résultat de ***Convergence en Loi***.

3.1.4 Variables de Poisson

Ces variables particulièrement utiles, découvertes au XIXème siècle par *Denis Siméon Poisson*, peuvent elles aussi être présentées en termes de *comptage* : imaginons, par exemple, qu'un gérant de station-service sache qu'il y a *en moyenne* $\lambda = 13,5$ clients se présentant dans sa station chaque matin entre 8h et 8h30 ; ce gérant pourra alors partir du principe que le nombre W de clients se présentant un matin donné entre 8h et 8h30 suit une loi de Poisson de paramètre $\lambda = 13,5$, ce qui signifie que

$$W(\Omega) = \mathbb{N} = \{0; 1; 2; \ldots\} \text{ et : } \forall k \in \mathbb{N}, \mathbb{P}\{W = k\} = e^{-\lambda}(\frac{\lambda^k}{k!})$$

Par exemple, la probabilité de voir trois clients exactement se présenter à la station-service le lendemain entre 8h et 8h30 vaut

$$\mathbb{P}\{W=3\} = e^{-\lambda}(\frac{\lambda^3}{3!}) = e^{-13,5} \times \frac{(13,5)^3}{6}$$

En fait, ces variables de Poisson doivent leur importance à un autre résultat fondamental de ***Convergence en Loi*** :
à $\lambda > 0$ fixé, si $n \to +\infty$ et si (p_n) est une suite tendant vers 0 comme $\frac{\lambda}{n}$, on a

$$\mathcal{B}(n; p_n) \approx \mathcal{P}(\lambda)$$

Plus précisément, en posant $p_n = \frac{\lambda}{n}$ on peut affirmer que

$$\mathcal{B}(n; p_n) \underset{n\to\infty}{\longrightarrow} \mathcal{P}(\lambda),$$

$\mathcal{P}(\lambda)$ désignant ici la loi de Poisson de paramètre λ. De façon plus détaillée :

$$\forall k \in \mathbb{N}, \quad \lim_{n\to\infty} \left\{ \binom{n}{k} p_n^k (1-p_n)^{n-k} \right\} = e^{-\lambda}(\frac{\lambda^k}{k!})$$

C'est cette convergence de la loi Binomiale $\mathcal{B}(n; p_n = \frac{\lambda}{n})$ vers la loi de Poisson $\mathcal{P}(\lambda)$ qui explique le grand succès des variables de Poisson, que l'on retrouve dans toutes sortes de contexte.
Par exemple, le comptage du nombre de premières connexions à une certaine page Internet durant une certaine fenêtre de temps pourra donner lieu à des calculs poissonniens (il y a un très grand nombre n d'internautes, mais, durant le temps considéré, chacun n'est susceptible de visiter ce site qu'avec une très petite probabilité p_n).
Voici un tout autre exemple : si l'on suppose que 10^5 bactéries sont présentes dans un litre d'eau et réparties de façon très diffuse à l'intérieur de ce volume, comment modéliser le comptage du nombre de bactéries présentes dans un mm^3 d'eau prélevé au hasard dans ce volume ? Pour ce comptage, il convient de considérer que le nombre W de bactéries obtenues suit une loi de Poisson de paramètre $\lambda = 10^5 \times 10^{-6} = 0,1$.
Une dernière remarque sur ces variables de Poisson : un seul paramètres λ permet de décrire leur loi ; λ est un nombre réel > 0 qui est de nature extensive (λ augmente avec l'amplitude de la fenêtre de temps, ou encore avec le volume d'eau).

3.1.5 Variables Géométriques

Ces variables sont souvent utilisées afin de modéliser un ***temps d'attente*** jusqu'à la première occurence d'un certain événement.
Supposons par exemple qu'Alice et Bob jouent indéfiniment à "Pile ou Face", et désignons par T le temps de première apparition d'un 'Pile'. Si la pièce utilisée produit 'Pile' avec probabilité p à chaque lancer, on aura

$$\forall k \geq 1, \quad \mathbb{P}\{T=k\} = (1-p)^{k-1} \cdot p$$

3.1.6 Variables Uniformes Discrètes

Ce sont des variables discrètes à valeurs dans un ensemble fini à n éléments, chaque valeur pouvant être prise avec la probabilité $\frac{1}{n}$.
Par exemple, si Alice et Bob s'amusent à lancer un dé icosaédrique (polyèdre régulier à vingt faces), la variable U décrivant le résultat du lancer peut être considérée comme une variable uniforme sur l'intervalle discret $[\![1;20]\!]$:

$$\forall k \in [\![1;20]\!], \quad \mathbb{P}\{T=k\} = \frac{1}{20}$$

3.2 Calculs d'Espérances, de Variances, de Moments

Comme nous venons de le voir, la loi d'une variable aléatoire discrète peut être considérée comme une répartition de masses de probabilité sur un certain ensemble de valeurs. Par exemple, considérer la loi binomiale $\mathcal{B}(N;p)$ revient à considérer un vecteur $(p_0;p_1;\ldots;p_k\ldots;p_n)$ de nombres positifs dont la somme vaut 1, avec

$$\forall k \in [\![0;n]\!], \; p_k = \mathbb{P}\{Y=k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

En termes physiques, le calcul de l'***Espérance*** d'une telle variable aléatoire correspond à la détermination du ***centre de gravité*** associé à cette répartition de masses ; pour une variable aléatoire discrète D à valeurs dans l'ensemble fini $V=\{v_1;v_2;\ldots;v_n\}$, on pose

$$\mathbb{E}[D] = \sum_{k=1}^{n} \mathbb{P}\{D=v_k\}\cdot v_k = \sum_{k=1}^{n} p_k v_k = p_1v_1 + p_2v_2 + \ldots + p_nv_n$$

En termes mathématiques, on retrouve la notion de ***barycentre***.
L'espérance $\mathbb{E}[D]$ d'une v.a. D est aussi appelée sa ***valeur moyenne***, ou encore son ***moment d'ordre*** 1.
Le ***moment d'ordre*** 2 d'une telle variable est ensuite défini comme l'espérance de son carré, et peut se calculer comme suit :

$$\mathbb{E}[D^2] = \sum_{k=1}^{n} \mathbb{P}\{D=v_k\}\cdot v_k^2 = \sum_{k=1}^{n} p_k v_k^2 = p_1v_1^2 + p_2v_2^2 + \ldots + p_nv_n^2$$

Plus généralement, le ***moment d'ordre*** m d'une telle variable est défini comme l'espérance de sa puissance d'ordre m et peut se calculer comme suit :

$$\mathbb{E}[D^m] = \sum_{k=1}^{n} \mathbb{P}\{D=v_k\}\cdot v_k^m = \sum_{k=1}^{n} p_k v_k^m = p_1v_1^m + p_2v_2^m + \ldots + p_nv_n^m$$

Dans la pratique, les deux premiers moments sont de loin les plus importants.
Le deuxième moment est toujours supérieur ou égal au carré du premier moment ($\mathbb{E}[D^2] \geq$

$\mathbb{E}[D]^2$), et en soustrayant le carré du premier moment au deuxième moment, on obtient la ***Variance*** de la variable considérée :

$$\mathrm{Var}(D) = \mathbb{E}[D^2] - \mathbb{E}[D]^2$$

Tout comme pour les variables statistiques, il y a en fait deux définitions équivalentes de la variance d'une telle variable D :

1. La variance de D vaut la différence entre le deuxième moment et le carré du premier moment :
$$\mathrm{Var}(D) = \mathbb{E}[D^2] - \mathbb{E}[D]^2$$
2. Elle vaut aussi la moyenne des carrés des écarts à la moyenne :
$$\mathrm{Var}(D) = \mathbb{E}\left[(D - \mathbb{E}[D])^2\right] = \sum_{k=1}^{n} p_k (v_k - \mu)^2,$$
où $\mu = \mathbb{E}[D]$ désigne la valeur moyenne de D.

En termes physiques, calculer une variance revient à déterminer un moment d'inertie pour la répartition de masses considérée.
Voyons tout de suite quelques exemples classiques de calculs d'espérance et de variance :

3.2.1 Variables de Bernoulli

Si X suit une loi de Bernoulli de paramètre p, on a

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}\{X = 0\} + 1 \cdot \mathbb{P}\{X = 1\} = p$$

et

$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}\{X = 0\} + 1^2 \cdot \mathbb{P}\{X = 1\} = p,$$

puis

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

Remarquons que $\mathrm{Var}(X) \geq 0$, l'égalité $\mathrm{Var}(X) = 0$ ne se produisant que pour des variables de Bernoulli "dégénérées", càd telles que $p = 0$ ou $p = 1$.

3.2.2 Variables Binomiales

Pour une variable binomiale Y de paramètres n et p, on obtient

$$\begin{array}{rcl} \mathbb{E}[Y] & = & 0 \cdot \mathbb{P}\{Y = 0\} + 1 \cdot \mathbb{P}\{Y = 1\} + \ldots + n \cdot \mathbb{P}\{Y = n\} \\ & = & \sum_{k=0}^{n} k \cdot \mathbb{P}\{Y = k\} \\ & = & \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ & = & np \end{array}$$

et

$$\begin{aligned}\mathbb{E}[Y^2] &= 0^2 \cdot \mathbb{P}\{Y=0\} + 1^2 \cdot \mathbb{P}\{Y=1\} + \ldots + n^2 \cdot \mathbb{P}\{Y=n\} \\ &= \sum_{k=0}^{n} k^2 \cdot \mathbb{P}\{Y=k\} \\ &= np(1-p) + n^2p^2,\end{aligned}$$

en sorte que

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = np(1-p)$$

Par exemple, si Y suit une loi $\mathcal{B}(n=10; p=0,5)$ et Y' une loi $\mathcal{B}(n'=20; p=0,25)$, Y et Y' ont même valeur moyenne :

$$\mathbb{E}[Y'] = \mathbb{E}[Y] = 5$$

(car $np = n'p' = 5$), mais la seconde variable est "plus dispersée" que la première :

$$\mathrm{Var}(Y') = n'p'(1-p') = \frac{15}{4} > \frac{5}{2} = np(1-p) = \mathrm{Var}(Y)$$

3.2.3 Variables Hypergéométriques

Pour une variable hypergéométrique Z de paramètres N, n et p, en supposant toujours que $n \leq \min(Np; N(1-p))$ on obtient

$$\begin{aligned}\mathbb{E}[Z] &= 0 \cdot \mathbb{P}\{Z=0\} + 1 \cdot \mathbb{P}\{Z=1\} + \ldots + n \cdot \mathbb{P}\{Z=n\} \\ &= \sum_{k=0}^{n} k \cdot \mathbb{P}\{Z=k\} \\ &= \sum_{k=0}^{n} k \cdot \frac{\binom{pN}{k}\binom{(1-p)N}{(n-k)}}{\binom{N}{n}} \\ &= np\end{aligned}$$

La valeur moyenne de Z est donc identique à celle d'une variable binomiale de paramètres n et p : le fait de procéder à des tirages *avec* ou *sans remise* n'a pas d'incidence sur la valeur moyenne de la variable de comptage associée.
En revanche, pour ce qui est du moment d'ordre 2 et de la variance, une différence surgit entre le cas hypergéométrique et le cas binomial ; on a en effet

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = np(1-p)\frac{N-n}{N-1} \leq np(1-p) = \mathrm{Var}(Y)$$

3.2.4 Variables de Poisson

Soit maintenant W une variable de Poisson de paramètre $\lambda > 0$.
Rappelons que $W(\Omega) = \mathbb{N}$ et que

$$\forall k \in \mathbb{N},\ \mathbb{P}\{W=k\} = e^{-\lambda}\left(\frac{\lambda^k}{k!}\right)$$

Formellement, l'évaluation de $\mathbb{E}[W]$ se fait donc en calculant la somme

$$\sum_{k=O}^{+\infty} k \cdot \mathbb{P}\{W=k\} = \sum_{k=O}^{+\infty} k \cdot e^{-\lambda}\left(\frac{\lambda^k}{k!}\right)$$

Le fait de manipuler une ***somme comportant une infinité de termes*** peut paraître gênant a priori. En fait, la valeur de la somme $\sum_{k=O}^{+\infty} k \cdot e^{-\lambda}(\frac{\lambda^k}{k!})$ peut être définie en toute rigueur ***comme limite de sommes finies*** ; après quelques calculs, il s'avère que

$$\begin{array}{rcl}
\mathbb{E}[W] & = & 0 \cdot \mathbb{P}\{W=0\} + 1 \cdot \mathbb{P}\{W=1\} + \ldots + k \cdot \mathbb{P}\{W=k\} + \ldots \\
& = & \sum_{k=0}^{\infty} k \cdot \mathbb{P}\{W=k\} \\
& = & \sum_{k=0}^{\infty} k \cdot e^{-\lambda}(\frac{\lambda^k}{k!}) \\
& = & \lim\limits_{K\to\infty} \sum_{k=0}^{K} k \cdot e^{-\lambda}(\frac{\lambda^k}{k!}) \\
& = & \lambda
\end{array}$$

L'évaluation de la variance de W passe par celle de son deuxième moment ; le même type de calculs donne pour ce deuxième moment :

$$\begin{array}{rcl}
\mathbb{E}[W^2] & = & 0^2 \cdot \mathbb{P}\{W=0\} + 1^2 \cdot \mathbb{P}\{W=1\} + \ldots + k^2 \cdot \mathbb{P}\{W=k\} + \ldots \\
& = & \sum_{k=0}^{\infty} k^2 \cdot \mathbb{P}\{W=k\} \\
& = & \sum_{k=0}^{\infty} k^2 \cdot e^{-\lambda}(\frac{\lambda^k}{k!}) \\
& = & \lim\limits_{K\to\infty} \sum_{k=0}^{K} k^2 \cdot e^{-\lambda}(\frac{\lambda^k}{k!}) \\
& = & \lambda^2 + \lambda,
\end{array}$$

en sorte que

$$\mathrm{Var}(W) = \mathbb{E}[W^2] - \mathbb{E}[W]^2 = \lambda$$

Les variables de Poisson ont donc ceci de particulier que leurs espérances coïncident avec leurs variances :

$$\mathbb{E}[W] = \mathrm{Var}(W) = \lambda$$

3.2.5 Variables Géométriques

Rappelons que pour une variable géométrique T de paramètre p :

$$T(\Omega) = \mathbb{N}^* \text{ et : } \forall k \geq 1, \quad \mathbb{P}\{T=k\} = (1-p)^{k-1} \cdot p$$

Tout comme pour les variables de Poisson, les calculs de $\mathbb{E}[T], \mathbb{E}[T^2]$ et $\mathrm{Var}(T)$ se feront en évaluant des sommes infinies obtenues comme limites de sommes finies.
On a plus précisément dans le contexte géométrique :

$$\begin{array}{rcl}
\mathbb{E}[T] & = & 1 \cdot \mathbb{P}\{T=1\} + 2 \cdot \mathbb{P}\{T=2\} + \ldots + k \cdot \mathbb{P}\{T=k\} + \ldots \\
& = & \sum_{k\geq 1} k \cdot \mathbb{P}\{T=k\} \\
& = & \sum_{k\geq 1} k \cdot (1-p)^{k-1} \cdot p \\
& = & \lim\limits_{K\to\infty} \sum_{k=1}^{K} k \cdot (1-p)^{k-1} \cdot p \\
& = & \frac{1}{p}
\end{array}$$

puis

$$\begin{array}{rcl}\mathbb{E}[T^2] & = & 1^2 \cdot \mathbb{P}\{T=1\} + 2^2 \cdot \mathbb{P}\{T=2\} + \ldots + k^2 \cdot \mathbb{P}\{T=k\} + \ldots \\ & = & \sum_{k\geq 1} k^2 \cdot \mathbb{P}\{T=k\} \\ & = & \sum_{k\geq 1} k^2 \cdot (1-p)^{k-1} \cdot p \\ & = & \lim\limits_{K\to\infty} \sum_{k=1}^{K} k^2 \cdot (1-p)^{k-1} \cdot p \\ & = & \frac{2}{p^2} - \frac{1}{p}, \end{array}$$

en sorte que

$$\mathrm{Var}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{1}{p^2} - \frac{1}{p}$$

3.2.6 Variables Uniformes Discrètes

Pour une variable aléatoire U distribuée uniformément sur l'intervalle discret $[\![1;n]\!]$, l'évaluation des deux premiers moments de U passe par la connaissance des valeurs de deux sommes finies classiques : la somme d'entiers consécutifs

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

et la somme de carrés d'entiers consécutifs

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

L'utilisation de ces deux valeurs de sommes conduit aux résultats suivants :

$$\mathbb{E}[U] = 1 \cdot \mathbb{P}\{U=1\} + 2 \cdot \mathbb{P}\{U=2\} + \ldots + n \cdot \mathbb{P}\{U=n\} = \sum_{k=1}^{n} k \cdot \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k = \frac{n+1}{2}$$

et

$$\mathbb{E}[U^2] = 1^2 \cdot \mathbb{P}\{U=1\} + 2^2 \cdot \mathbb{P}\{U=2\} + \ldots + n^2 \cdot \mathbb{P}\{U=n\} = \sum_{k=1}^{n} k^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k^2 = \frac{(n+1)(2n+1)}{6}$$

On a donc :

$$\mathrm{Var}(U) = \mathbb{E}[U^2] - \mathbb{E}[U]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

Les résultats importants énoncés dans ce paragraphe sont récapitulés dans le tableau qui suit.

3.2.7 Tableau Récapitulatif

	$V(\Omega) =$	Loi de V	$\mathbb{E}[V]$	$\mathrm{Var}(V)$
$\mathcal{B}(n;p)$	$[\![0;n]\!]$	$\mathbb{P}\{V=k\} = \binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
$\mathcal{H}(N;n;p)$	$[\![0;n]\!]$	$\mathbb{P}\{V=k\} = \frac{\binom{pN}{k}\binom{(1-p)N}{(n-k)}}{\binom{N}{n}}$	np	$np(1-p)\frac{N-n}{N-1}$
$\mathcal{P}(\lambda)$	$\mathbb{N}$	$\mathbb{P}\{V=k\} = e^{-\lambda}(\frac{\lambda^k}{k!})$	λ	λ
$\mathcal{G}(p)$	$\mathbb{N}^*$	$\mathbb{P}\{V=k\} = (1-p)^{k-1}\cdot p$	$\frac{1}{p}$	$\frac{1}{p^2}-\frac{1}{p}$
$\mathcal{U}([\![1;n]\!])$	$[\![1;n]\!]$	$\mathbb{P}\{V=k\} = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$

3.3 Variables Aléatoires Indépendantes

Soient $X, Y : \Omega \longrightarrow \mathbb{R}$ deux variables aléatoires définies sur le même espace de probabilité $(\Omega;\mathcal{F};\mathbb{P})$. X et Y seront dites ***indépendantes*** dès lors que tous les événements du type

$$(X \in I) = \{\omega \in \Omega | X \in I\} \text{ et } (Y \in J) = \{\omega \in \Omega | Y \in J\}$$

sont indépendants, quels que soient les intervalles $I, J \subset \mathbb{R}$, ce qui signifie que

$$\mathbb{P}\left\{(X \in I) \text{ et } (Y \in J)\right\} = \mathbb{P}(X \in I)\cdot\mathbb{P}(Y \in J)$$

Intuitivement, la donnée d'une information relative au comportement de X, du type $'X \in I'$, ne nous renseigne alors en rien sur le comportement de Y. Et vice-versa.
Dans le contexte des variables discrètes, il suffit de considérer les cas où I et J sont des singletons. Etablir l'indépendance des variables X et Y revient alors à montrer que

$$\forall a \in X(\Omega), \forall b \in Y(\Omega), \quad \mathbb{P}\left\{(X=a) \text{ et } (Y=b)\right\} = \mathbb{P}(X=a)\cdot\mathbb{P}(Y=b)$$

Plus généralement, trois variables discrètes $X, Y, Z : \Omega \longrightarrow \mathbb{R}$ définies sur un même espace de probabilité seront dites indépendantes (ou indépendantes dans leur ensemble) dès lors que

$$\begin{aligned}
\forall a \in X(\Omega), \forall b \in Y(\Omega), \forall c \in Z(\Omega), \quad \mathbb{P}\left\{(X=a)\cap(Y=b)\right\} &= \mathbb{P}(X=a)\cdot\mathbb{P}(Y=b) \\
\mathbb{P}\left\{(X=a)\cap(Z=c)\right\} &= \mathbb{P}(X=a)\cdot\mathbb{P}(Z=c) \\
\mathbb{P}\left\{(Y=b)\cap(Z=c)\right\} &= \mathbb{P}(Y=b)\cdot\mathbb{P}(Z=c) \\
\text{et } \mathbb{P}\left\{(X=a)\cap(Y=b)\cap(Z=c)\right\} &= \mathbb{P}(X=a)\cdot\mathbb{P}(Y=b)\cdot\mathbb{P}(Z=c)
\end{aligned}$$

Intuitivement, chaque variable *"ignore tout du comportement des autres variables"*. On a alors par exemple :

$$\forall a \in X(\Omega), \forall b \in Y(\Omega), \forall c \in Z(\Omega), \quad \mathbb{P}_{(Y=b)}(X=a) = \mathbb{P}_{(Z=c)}(X=a) = \mathbb{P}(X=a)$$

3.4 Espérances et Variances de Sommes

Terminons ce chapitre par une liste de propriétés remarquables de l'Espérance et de la Variance, les plus importantes d'entre elles portant sur ***l'Espérance et la Variance d'une somme de variables aléatoires***. Dans tout ce paragraphe, X et Y sont deux variables aléatoires définies sur un même espace de probabilité $(\Omega;\mathcal{F};\mathbb{P})$ et pour lesquelles $\mathbb{E}[X]$ et $\mathbb{E}[Y]$ existent.

1. L'Espérance est ***monotone*** : si X et Y et vérifient
$$\forall \omega \in \Omega, \quad X(\omega) \leq Y(\omega)$$
on aura
$$\mathbb{E}[X] \leq \mathbb{E}[Y]$$
Par exemple, si un dé à 6 faces est lancé 18 fois et que X compte le nombre d'apparitions de '4' tandis que Y compte le nombre d'apparitions d'un chiffre pair, on a
$$\mathbb{E}[X] = 18 \times \frac{1}{6} = 3 \leq 9 = 18 \times \frac{1}{2} = \mathbb{E}[Y]$$

2. Par voie de conséquence, si $X(\Omega) \subset [a;b]$, on aura nécessairement $a \leq X \leq b$.
Par exemple, l'espérance d'une variable binomiale de paramètres n et p est toujours située entre 0 et n, quel que soit p.

3. L'Espérance est ***linéaire*** : si α et β sont deux réels quelconques, on aura toujours
$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$$
Cette propriété peut aussi être énoncée en deux temps : on aura toujours
$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \text{ ainsi que } \mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$$
Bien entendu, la linéarité de l'Espérance vaut tout aussi bien pour trois v.a. ou plus. Par exemple, si Z est une troisième v.a. définie sur le même espace Ω et admettant elle aussi une espérance, on pourra utiliser le fait que
$$\mathbb{E}[\alpha X + \beta Y + \gamma Z] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y] + \gamma\mathbb{E}[Z]$$

Venons-en à quelques propriétés remarquables de la Variance. Pour énoncer ces propriétés convenablement, il nous faut supposer que les variables aléatoires $X, Y : \Omega \longrightarrow \mathbb{R}$ admettent des moments d'ordre deux, mais aussi définir la notion de ***Covariance*** de ces deux variables :
$$\boldsymbol{Cov}(X;Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$
Ici encore, tout comme dans le contexte des variables statistiques, la notion de Covariance généralise celle de Variance au sens où
$$\boldsymbol{Cov}(X;X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X)$$
Voici donc quelques propriétés remarquables de la Variance et de la Covariance :

1. ***Positivité :*** on aura toujours
$$\mathrm{Var}(X) \geq 0$$
l'égalité $\mathrm{Var}(X) = 0$ ne se produisant que si la variable X prend une seule et même valeur avec probabilité 1.

2. La Variance est ***Quadratique***, au sens où
$$\forall \alpha \in \mathbb{R}, \quad \mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$$

3. ***Variance d'une somme, d'une combinaison linéaire :*** on aura toujours
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\mathrm{Cov}(X;Y) + \mathrm{Var}(Y),$$
et, plus généralement :
$$\forall \alpha, \beta \in \mathbb{R}, \quad \mathrm{Var}(\alpha X + \beta Y) = \alpha^2 \mathrm{Var}(X) + 2\alpha\beta \mathrm{Cov}(X;Y) + \beta^2 \mathrm{Var}(Y)$$

4. ***Variance d'une somme de variables indépendantes :*** si les variables X et Y sont indépendantes, on aura $\mathrm{Cov}(X;Y) = 0$.
Par voie de conséquence, l'***indépendance*** des variables X et Y entraîne une certaine ***additivité de la Variance*** :
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$
Attention, il faut voir cette additivité de la Variance comme une conséquence de l'indépendance de X et Y, elle n'est pas du tout vraie en général ! Ainsi a-t-on, dès lors que $\mathrm{Var}(X) > 0$:
$$\mathrm{Var}(X+X) = \mathrm{Var}(2X) = 4\mathrm{Var}(X) \neq \mathrm{Var}(X) + \mathrm{Var}(X) = 2\mathrm{Var}(X)$$
Bien entendu, cette règle d'additivité vaut aussi pour la Variance d'une somme de plus de deux variables indépendantes : si $X_1, X_2, \ldots, X_n : \Omega \longrightarrow \mathbb{R}$ sont des v.a. ***indépendantes*** et admettant toutes des moments d'ordre 2, on pourra écrire
$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n),$$
ou encore, de manière plus compacte :
$$\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathrm{Var}(X_k)$$

3.5 Exercices d'Application

I Quelques lois classiques :

Soient X une variable $\mathcal{B}(8; 1/4)$, Y une variable $\mathcal{P}(2)$ et Z une variable $\mathcal{G}(1/4)$.

1. Calculer les prob. des événements

$$\{X \geq 4\}, \{1 \leq X < 4\}, \{Y \geq 4\}, \{1 \leq Y < 4\}, \{Z \geq 4\}, \{1 \leq Z < 4\}.$$

2. Quels sont les ***modes*** (valeurs les plus probables) de ces trois variables ?

II Quelques variables de comptage :

1. Une famille de dauphins est composée de six femelles et de quatre mâles. Un groupe de quatre dauphins est constitué au hasard dans cette famille ; soit X la variable comptant le nombre de femelles que l'on peut observer dans ce groupe. Déterminer la loi de X puis donner son mode, son espérance et sa variance.
2. Un restaurant comporte 50 places. La probabilité pour qu'une personne ayant réservé ne vienne pas est de 20%. Un jour, le patron accepte de prendre en compte 52 réservations ; quelle est la probabilité qu'il se retrouve dans une situation embarrassante ?
3. Dans le but d'estimer la taille d'une certaine population animale, une équipe d'écologues capture r animaux de l'espèce considérée et marque ces r animaux avant de les relâcher et de leur laisser le temps de se disperser dans le territoire étudié. Ils procèdent ensuite à une nouvelle capture de n animaux de la même espèce, avec $n > r$, et dans un temps suffisamment rapproché pour que la structure de cette population n'aie pas connu de changement notable. Soit Z le nombre d'animaux marqués parmi les n animaux de la deuxième capture. Quelle est la loi de Z ? Quel est son mode ? Quelle est la valeur la plus vraisemblable de la taille N de cette population d'animaux ?
4. Le nombre Y d'oeufs pondus par une tortue au cours d'une ponte suit une loi $\mathcal{P}(12)$; on suppose en outre que chaque oeuf pondu a une prob. $p = 0,75$ d'arriver à éclosion. Quelle est la loi du nombre $\tilde{Y}$ de bébés tortue obtenus à l'issue d'une ponte ?
5. Un bureau de poste comporte deux guichets : le premier est réservé aux services bancaires, le second au courrier. On suppose que le nombre de clients se présentant au premier guichet entre $8H$ et $9H$ suit une loi de Poisson de paramètre $\lambda_1 = 9$, tandis que pour le second guichet ce nombre est une variable de Poisson de paramètre $\lambda_2 = 13$. On suppose en outre que ces deux variables sont indépendantes. Soit W la variable comptant le nombre total de clients du bureau. Quelle est la loi de W?

III Virus et Poisson :

Un virus informatique se présente sous deux formes: A et B, et se transmet de manières distinctes pour les deux formes, si bien que l'on peut supposer que le nombre de postes infestés par la forme A est indépendant du nombre de postes infestés par la forme B. Un poste ne peut pas être atteint simultanément par les 2 formes de virus. On note X et Y les nombres (aléatoires) de postes infestés par A et B respectivement, et on suppose que X et Y suivent des lois de Poisson de paramètres respectifs α et β.

1. Quel est le nombre moyen de postes infectés par la forme A du virus? par la forme B? par l'une des 2 formes?

2. Que représente $X+Y$? Quelle est la valeur moyenne de $X+Y$? Quelle est la probabilité que $X+Y$ vale 0?

3. Montrer que $X+Y$ suit une loi de Poisson de paramètre $\alpha+\beta$.

4. ⊞ Effectuer des simulations pour "vérifier" le résultat de la question 3).

IV Un guichet poissonien :

Le nombre de personnes se présentant à un guichet administratif à une heure donnée est aléatoire et distribué suivant une loi de Poisson de paramètre λ (λ est une constante > 0).

1. Etant établi que la probabilité que personne ne se présente à cette heure est de 5%, déterminer la valeur de λ.

2. Quel est le nombre moyen de personnes se présentant ?

3. Combien faut-il prévoir de guichets si on veut une garantie de 95% que personne n'attende.

V Approximations Poissoniennes :

1. Le Département *R&D* de la NASA doit donner son avis sur un projet de navette spatiale. L'équipement électronique de la navette comporte 2 millions de pièces distinctes, et chacune de ces pièces a une probabilité $p = 10^{-7}$ de tomber en panne, indépendamment du fonctionnement des autres pièces. Le dysfonctionnement d'une seule de ces pièces suffirait à entraîner une explosion immédiate de la navette. Que pensez-vous de ce projet ?
2. Dans un pays lointain, un réservoir de 2'000 litres d'eau contient des insectes exotiques, avec en moyenne de 2 insectes par litre. On admet qu'il est dangereux d'avaler au moins 8 insectes. Un touriste assoiffé boit un litre d'eau de ce réservoir ; a-t-il réellement mis sa vie en danger ?

VI Couples de variables aléatoires :

1. Bob ne supporte pas les chats et Alice déteste les chiens. Bob n'élève pas plus d'un chien et Alice pas plus d'un chat. La probabilité que Bob possède un chien vaut $0,2$; si Bob n'a pas de chien, Alice possède un chat avec probabilité $0,1$, tandis que si Bob a un chien cette probabilité monte à $0,5$. Soit X_1 le nombre de chiens de Bob et X_2 le nombre de chats d'Alice. Donner la loi du couple $(X_1; X_2)$ de variables aléatoires, puis la loi marginale de X_2, puis enfin la loi de $X = X_1 + X_2$ (nombre total d'animaux).
2. A l'oral d'un concours, un candidat doit composer sur trois sujets tirés au sort parmi huit, dont 3 sujets d'économie, 2 sujets de droit et 3 de gestion. Soit Y_1 le nombre de sujets de droit sortis et Y_2 le nombre de sujets d'économie sortis. Donner la loi du couple $(Y_1; Y_2)$, puis les lois marginales de Y_1 et Y_2, puis enfin la loi conditionnelle de Y_2 sachant qu'aucun sujet d'économie n'a été choisi.

VII Sur la Technique de "Pooling" :

On propose une méthode économique pour analyser des échantillons sanguins afin de détecter une certaine maladie. Pour cela, on fait subir à chaque individu de la population sous surveillance un prélèvement qui est analysé de la façon suivante:

- sur chaque prélèvement individuel, on retire une dose;

- on regroupe les doses 10 par 10, ce qui constitue un lot; chaque lot est analysé;

- si le résultat du lot est négatif, c'est que la maladie n'est présente dans aucune des 10 doses;

- si le résultat du lot est positif, c'est que la maladie est présente dans au moins l'une des doses; on procède alors à l'analyse de chaque dose individuelle.

On prélève 120 échantillons chaque jour. On suppose que la proportion de malades dans la population est 2% et que les états des individus (malade ou pas) sont indépendants.

1. Pour un lot de 10 prélèvements donné, quelle est la probabilité que le résultat de l'analyse soit négatif?
2. Soit X le nombre de lots positifs obtenus au cours d'une journée. Quelle est la loi de X? Déterminer son espérance.
3. Soit Y le nombre total d'analyses à faire dans une journée. Exprimer Y en fonction de X, calculer l'espérance de Y ainsi que la probabilité que $Y > 120$.
4. Comparer la méthode d'analyse proposée avec la méthode simple qui consiste à analyser séparément chaque prélèvement.
5. Et si on avait regroupé les doses par lots de 8 ? Par lots de 12 ?

VIII Panini Forever :

Votre petit-cousin collectionne les images des joueurs de la coupe du monde, que l'on trouve exclusivement dans les tablettes de chocolat de la marque *"Scrounch"*. Il y a en tout N portraits de footballeurs à collectionner, et ces N images sont réparties équitablement, à raison d'une seule image par tablette de chocolat. Soit X_r le nombre de tablettes achetées

jusqu'à obtention de r images distinctes, puis T_n le nombre de tablettes qu'il faut acheter pour tomber sur une nouvelle image sachant que l'on possède déjà $(n-1)$ images distinctes $(T_1 = 1)$.

1. Donner la loi de T_2, puis celle de T_n.
2. En déduire la valeur de $\mathbb{E}[X_r]$ $(1 \leq r \leq N)$ puis en particulier celle de $\mathbb{E}[X_N]$.
3. On admet que les v.a. T_n $(n \geq 1)$ sont indépendantes ; calculer $\mathrm{Var}(X_N)$.
4. Votre petit-cousin s'intéresse en fait uniquement aux joueurs de l'équipe luxembourgeoise. On note Y_k le nombre de tablettes achetées jusqu'à obtention de ces k joueurs. Evaluer $\mathbb{E}[Y_k]$ et $\mathrm{Var}(Y_k)$.

IX Simulations d'une loi non-uniforme dans $\{1, \ldots, N\}$ ⊞ :

Construire, en faisant appel à la fonction `rand`, une fonction que vous appelerez `nonunif` qui prend comme variable d'entrée un vecteur p de longueur N quelconque et fournit en sortie,
- soit la réponse "`p n'est pas une probabilité`" si la somme des coordonnées de p ne vaut pas 1,
- soit, si la somme des coordonnées de p vaut 1, un entier aléatoire entre 1 en N de telle sorte que la probabilité d'obtenir i soit égale à p_i lorsque $p = (p_1, p_2, \ldots, p_N)$.

X Convergences en loi ⊞ :

1. *Hypergéométrique→Binomiale* :
   (a) Pour différentes valeurs (bien choisies) de N, n et p, comparer les histogrammes des lois $\mathcal{H}(N; n; p)$ et $\mathcal{B}(n; p)$ (on pourra utiliser *Excel* ou encore *Scilab*).
   (b) Utiliser la distance euclidienne entre vecteurs de probabilités $\boldsymbol{p}$ et $\boldsymbol{p'}$, donnée par

$$d(\boldsymbol{p}; \boldsymbol{p'}) = \sqrt{(p'_0 - p_0)^2 + (p'_1 - p_1)^2 + \ldots + (p'_n - p_n)^2} = \sqrt{\sum_{i=0}^{n} (p'_i - p_i)^2}$$

   pour parvenir à une appréciation plus précise de la convergence annoncée en cours.
   (c) Remplacer la distance euclidienne de la question précédente par d_∞, définie par

$$d_\infty(\boldsymbol{p}; \boldsymbol{p'}) = \max_{0 \leq i \leq n} |p'_i - p_i|$$

   Que constatez-vous ?
2. *Binomiale→Poisson* :
   reprendre les questions précédentes afin de vérifier la validité de la convergence *Binomiale→Poisson* annoncée en cours.

XI Encore des dés :

On lance deux dés à six faces équilibrés, puis on note X le résultat produit par le premier dé, Y celui produit par le second dé et $S = X + Y$ la somme de ces deux résultats.

QCM1 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}(X) = 3$

B $\mathbb{E}(X) = 3,5$

C $\mathbb{E}(X^2) = 15,17$

D $\text{Var}(X) = 2,92$

E $\text{Var}(X) = 6,17$

QCM2 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(X = 1, Y = 3) = 0,028$

B $\mathbb{P}(S = 4) = 0,083$

C $\mathbb{P}(S = 4) = 0,111$

D $\mathbb{P}(X = 1|S = 4) = 0,25$

E $\mathbb{P}(X = 1|S = 4) = 0,33$

QCM3 : Parmi les propositions suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A Les v.a. X et Y sont indépendantes conditionnellement à S.

B $\mathbb{P}(S = 7|\, X = 1) = 0,17$

C $\mathbb{P}(S = 7|\, X = 1) = 0,33$

D $\mathbb{P}(S = 2) = \mathbb{P}(S = 12)$

E X suit une loi binomiale.

XII Comparaisons de traitements :

Face à une maladie gravissime, constamment mortelle après cinq ans au maximum, on peut réaliser un traitement médical ou un traitement chirurgical. Les probabilités associées aux différentes durées de vie après traitement sont indiquées ci-dessous :

Durée de vie (en années) / Traitement	1	2	3	4	5
Médical	0,2	0,3	0,3	0,2	0
Chirurgical	0,6	0,1	0,1	0,1	0,1

On note respectivement D_C et D_M les variables de durée de vie après traitement chirurgical et après traitement médical.

QCM4 : Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}(D_M) = 2$

B $\mathbb{E}(D_M) = 3$

C $\mathbb{E}(D_C) = 2,5$

D $\mathbb{E}(D_C) = 3$

E $\mathrm{Var}(D_C) > \mathrm{Var}(D_M)$

QCM5 : Pour deux patients A et B, on définit des fonctions d'utilité U_A et U_B associées à ces durées de vie. Ces fonctions reflètent les préférences des patients, et peuvent donc différer d'un patient à l'autre.

Dans le cas du patient A, la fonction U_A est telle que $U_A(D) = D$ pour toute valeur $D \in [\,[1;5]\,]$; en revanche, dans le cas du patient B on a

$$U_B(D) = \begin{cases} -10 & \text{si } D = 1 \\ 2 & \text{si } 2 \leq D \leq 5 \end{cases}$$

Pour un patient donné, le meilleur traitement est celui dont l'utilité est d'espérance mathématique maximale, les durées de vie étant elles-mêmes des variables aléatoires.

Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}\left[U_A(D_M)\right] = \mathbb{E}(D_M)$

B $\mathbb{E}\left[U_B(D_C)\right] = \mathbb{E}(D_C)$

C $\mathbb{E}\left[U_B(D_M)\right] = 2,6$

D Pour le patient A, le traitement médical est le meilleur.

E Pour le patient B, le traitement chirurgical est le meilleur.

XIII Etude de deux variables aléatoires :

La distribution conjointe de deux variables aléatoires X et Y est décrite ci-dessous :

X \ Y	0	1	2
0	0,15	0,2	0,25
1	0,2	0,1	0,1

QCM6 : Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(X = 0) = 0,4$

B $\mathbb{P}(X = 0) = 0,6$

C $\mathbb{P}(Y = 1) = 0,3$

D $\mathbb{P}(Y > 0) = 0,4$

E X suit une loi de Bernoulli.

QCM7 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}(X) = 0,4$

B $\mathbb{E}(X) = 0,6$

C $\mathbb{E}(Y^2) = 0,6$

D $\mathbb{E}(Y^3) = 1,7$

E $\mathbb{E}(Y^4) = 5,9$

QCM8 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathrm{Var}(Y) = 0,7$

B $\mathbb{E}(XY) = -0,3$

C $\mathbb{E}(XY) = 0,3$

D $\mathrm{Cov}(X, Y) = -0,1$

E $\mathrm{Cov}(X, Y) = 0,1$

QCM9 : On note $\rho(X, Y)$ le coefficient de corrélation des variables X et Y (quotient de leur covariance par le produit de leurs écarts-types).Peut-on affirmer que

A $\rho(X, Y) = -0,73$?

B $\rho(X, Y) = -0,24$?

C $\rho(X, Y) = 0,73$?

D les v.a. X et Y sont indépendantes ?

E les v.a. X et Y ne sont pas indépendantes ?

QCM10 : On définit les deux nouvelles variables X et Y par $U = 2X + 3$ et $V = 5Y$. Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}(U) = 3,8$
B $\mathbb{E}(U) = 4,2$
C $\mathbb{E}(V) = 4$
D $\mathbb{E}(V) = 5$
E $\mathbb{E}(V) = 6$

QCM11 : peut-on affirmer que

A $\mathbb{E}(UV) = 14$?
B $\text{Cov}(U, V) = 0$?
C $\text{Cov}(U, V) = -1$?
D $\rho(U, V) = 0$?
E $\rho(U, V) = -0,24$?

QCM12 : on considère les variables $S = X + Y$ et $T = XY$. Peut-on affirmer que

A $\mathbb{P}(S = 0,\, T = 0) = 0,12$?
B $\mathbb{P}(S = 0,\, T = 0) = 0,15$?
C $\mathbb{P}(S = 0,\, T = 0) = 0,18$?
D $\mathbb{P}(S = 0,\, T > 0) = 0$?
E $\mathbb{P}(S = 0,\, T > 0) = 0,15$?

QCM13 : Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(S = 1) = 0,2$
B $\mathbb{P}(S = 2) = 0,4$
C $\mathbb{E}(S) = 1,1$
D $\mathbb{E}(S) = 1,4$
E $\mathbb{E}(T) = 0,3$

QCM14 : Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{E}(ST) = 0,2$
B $\mathbb{E}(ST) = 0,8$
C $\text{Cov}(S, T) = 0,38$
D $\text{Cov}(S, T) = 0,8$

E Les v.a. S et T sont indépendantes.

QCM15 : Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(S = 2 | X = 0) = 0,42$
B $\mathbb{P}(S = 2 | X = 0) = 0,58$
C $\mathbb{P}(X = 1 | T = 2) = 0,5$
D $\mathbb{P}(S > T) = 0,2$
E $\mathbb{P}(S > T) = 0,85$

XIV Facteurs de risque :

Une maladie M a cinq facteurs de risque F_1, F_2, F_3, F_4, F_5. La présence ou l'absence de chacun de ces facteurs de risque est modélisée par une loi de Bernoulli ; on suppose que la probabilité de présence de chaque facteur est égale à $0,3$, et que ces présences ou absences des facteurs sont indépendantes. Soit N le nombre de facteurs présents en tout chez un individu donné. **QCM16 :** Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A N suit une loi équiprobable.
B N suit une loi binomiale.
C N suit une loi de Poisson.
D $\mathbb{P}(N = 1) = 0,2$
E $\mathbb{P}(N = 1) = 0,36$

QCM17 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(N > 0) = 0,800$
B $\mathbb{P}(N > 0) = 0,832$
C $\mathbb{P}(N = 2) = 0,309$
D $\mathbb{P}(N = 3) = 0,132$
E $\mathbb{P}(N = 4) = 0,084$

XV Gardes de nuit :

Un Centre Médical embauche un jeune médecin pour faire des gardes de nuit. Le nombre de patients appellant le patient chaque nuit suit une loi de Poisson de paramètre $\lambda = 2$. Le Centre rémunère le médecin de la manière suivante : pour chaque appel, le médecin reçoit 30 Euros d'honoraires, mais s'il reçoit 0 ou 1 appel seulement, le Centre complète ces honoraires pour qu'il ait une rémunération minimale garantie de 50 Euros au total.
On note N le nombre d'appels durant une nuit de garde, H les honoraires du médecin à l'issue de cette nuit de garde (sans complément de rémunération) et T sa rémunération totale, comportant un éventuel complément.

QCM18 : Parmi les égalités suivantes, laquelle (ou lesquelles) est (sont) correctes ?

A $\mathbb{P}(N = 0) = 0,135$

B $\mathbb{P}(N = 1) = 0,271$

C $\mathbb{P}(N > 1) = 0,694$

D $\mathbb{E}(H) = 15$ Euros

E $\mathbb{E}(H) = 60$ Euros

QCM19 : Parmi les identités suivantes, laquelle (ou lesquelles) est (sont) exactes ?

A $\text{Var}(H) = 60$ Euros^2

B $\text{Var}(H) = 1800$ Euros^2

C $\mathbb{E}(T) = 12,17$ Euros

D $\mathbb{E}(T) = 50$ Euros

E $\mathbb{E}(T) = 72,17$ Euros

QCM20 : Les frais pour le Centre Médical sont de 10 Euros par nuit de garde, augmentés de la rémunération complémentaire éventuelle du médecin, soit 50 Euros s'il n'y a pas d'appel ou 20 Euros s'il y a un et un seul appel. Sa recette est constituée par 10 Euros reçus par appel. Le bénéfice par garde, noté B, est la différence entre la recette et les frais. Si l'espérance du bénéfice est nulle, le budget du Centre est *en équilibre* ; ce budget est dit *positif* si l'espérance du bénéfice est strictement positive, *négatif* si l'espérance du bénéfice est strictement négative.
Parmi les affirmations suivantes, laquelle (ou lesquelles) est (sont) exactes ?

A Le budget du centre est positif.

B Le budget du centre est négatif.

C Pour obtenir un budget en équilibre, il faudrait fixer la rémunération minimale garantie à 45 Euros (en arrondissant à l'Euro supérieur).

D Pour obtenir un budget en équilibre, il faudrait fixer la rémunération minimale garantie à 55 Euros (en arrondissant à l'Euro supérieur).

E Pour obtenir un budget en équilibre, il faudrait fixer la rémunération minimale garantie à 60 Euros (en arrondissant à l'Euro supérieur).

4 VARIABLES ALEATOIRES CONTINUES

4.1 Variables continues

Une variable aléatoire $X : \Omega \longrightarrow \mathbb{R}$ sera dite ***continue*** lorsque sa probabilité de prendre exactement une valeur donnée $v \in \mathbb{R}$ vaut toujours 0 :

$$\forall v \in \mathbb{R}, \quad \mathbb{P}\{X = v\} = 0$$

Pour étudier la loi ou le comportement d'une telle variable X, on s'intéressera non plus à des événements du type $\{X = v\}$, mais à la probabilité $\mathbb{P}\{v \leq X \leq w\}$ de voir X *"atterrir dans un intervalle donné"* $[v; w]$.
Pour évaluer de telles probabilités, on fera usage d'une fonction $f_X : I \longrightarrow \mathbb{R}$ définie sur l'intervalle $I = X(\Omega)$ et telle que

$$\begin{cases} f_X \geq 0 & (f_X \text{ est à valeurs positives ou nulles}) \\ \int_I f_X(x)dx = 1 & (f_X \text{ représente une masse de probabilité}) \end{cases}$$

f_X est la ***fonction de densité*** de la variable aléatoire X, et l'on aura pour tout sous-intervalle $[v; w]$ de $I = X(\Omega)$:

$$\mathbb{P}\{v \leq X \leq w\} = \int_v^w f(x)dx$$

On pourra aussi prolonger la densité f_X en la définissant aussi sur $(\mathbb{R} \setminus I)$ par

$$\forall x \in (\mathbb{R} \setminus I), \quad f_X(x) = 0$$

La ***fonction de répartition*** associée à cette fonction de densité est alors la fonction continue et croissante $F_X : \mathbb{R} \longrightarrow \mathbb{R}$ définie par

$$\forall x \in \mathbb{R}, \quad F_X(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{X \in]-\infty; x]\} = \int_{-\infty}^x f_X(t)dt$$

Spécifier la loi d'une variable aléatoire continue X revient donc à donner sa fonction de densité, ou encore sa fonction de répartition.
Pour conclure cette introduction aux variables aléatoires continues, soulignons que

1. L'évaluation d'une ***intégrale sur un intervalle non-borné*** tel que $]-\infty; x]$ se fait par un ***passage à la limite*** :

$$\int_{-\infty}^x f_X(t)dt = \lim_{v \to -\infty} \int_v^x f_X(t)dt$$

2. On aura toujours : $\lim\limits_{x \to -\infty} F_X(x) = 0$, $\lim\limits_{x \to +\infty} F_X(x) = 1$.

3. La fonction de répartition $F_X : \mathbb{R} \longrightarrow \mathbb{R}$ d'une variable aléatoire *continue* X est toujours *continue et croissante* sur $\mathbb{R}$.

4. Dans tous les exemples que nous verrons, la densité f_X sera elle-même continue sur l'intervalle I, en sorte que
$$\forall x \in I, \quad F'_X(x) = f_X(x)$$

Venons-en aux exemples les plus utiles de v.a. continues.

4.1.1 Variables uniformes

On dira que la variable continue $X : \Omega \longrightarrow \mathbb{R}$ ***suit une loi uniforme sur l'intervalle*** $[a; b]$ si sa fonction de densité f_X est donnée par

$$\forall x \in \mathbb{R}, \quad f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{sinon} \end{cases}$$

Par voie de conséquence, la fonction de répartition F_X d'une telle variable est donnée par

$$\forall x \in \mathbb{R}, \quad F_X(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } x \geq b \end{cases}$$

Exemple d'utilisation : Un astronome sait que la distance X séparant son laboratoire d'un certain satellite varie uniformément de quinze milliers à trente milliers de Km. Quelle est la probabilité qu'il enregistre une distance moindre que $22'530$ Km entre son Laboratoire et le Satellite, un matin donné ?
En exprimant X en milliers de Km, il vient

$$\mathbb{P}\{X \leq 22,53\} = F_X(22,53) = \mathbb{P}\{15 \leq X \leq 22,53\} = \frac{22,53-15}{30-15} = \frac{7,53}{15} = 0,502,$$

la probabilité recherchée vaut donc $50,2\%$. Pour l'événement $\{18 \leq X \leq 22,53\}$, on obtient la probabilité

$$\mathbb{P}\{18 \leq X \leq 22,53\} = \mathbb{P}\{X \leq 22,53\} - \mathbb{P}\{X \leq 18\} = F_X(22,53) - F_X(18) = \frac{22,53-18}{30-15} = \frac{4,53}{15} = 0,302.$$

4.1.2 Variables exponentielles

On dira que la variable continue $T : \Omega \longrightarrow]0; +\infty[$ suit une ***loi exponentielle de paramètre*** $\lambda > 0$ si sa fonction de densité f_T est donnée par

$$\forall x \in \mathbb{R}, \quad f_T(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

Par voie de conséquence, la fonction de répartition F_T d'une telle variable est donnée par

$$\forall x \in \mathbb{R}, \quad F_T(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

Dans la pratique, ces variables sont utilisées pour modéliser un temps d'attente, en lien avec les variables de Poisson.

Exemple d'utilisation : Un gérant de magasin sait que les clients affluent toute la journée dans son commerce au rythme de $\lambda = 12$ clients par heure en moyenne ; plus précisément, il part du principe que le nombre de clients venant durant une heure donnée de la journée suit une loi de Poisson de paramètre $\lambda = 12$. Ouvrant les portes du magasin un matin, il désigne par T le temps d'attente de l'arrivée du premier client.

Dans ces conditions, le temps d'attente T suit une loi exponentielle de paramètre $\lambda = 12$; par exemple, la probabilité qu'il faille attendre plus de 30mn jusqu'à l'arrivée du premier client vaut

$$\mathbb{P}\{T > 0,5\} = 1 - \mathbb{P}\{T \leq 0,5\} = 1 - F_X(0,5) = 1 - (1 - e^{-12\times 0,5}) = e^{-6} \approx 2,5 \cdot 10^{-3}$$

cette probabilité ne vaut donc que $0,25\%$. Pour l'événement $\{0,25 \leq T \leq 0,5\}$, correspondant à un temps d'attente situé entre 15 et 30mn, on obtient

$$\mathbb{P}\{0,25 \leq T \leq 0,5\} = \mathbb{P}\{X \leq 0,5\} - \mathbb{P}\{X < 0,25\} = F_X(0,5) - F_X(0,25) = e^{-12\times 0,25} - e^{-12\times 0,5} \approx 0,0473$$

Il est à remarquer que

- Le réglage du paramètre λ se fait en fonction du choix d'une unité de mesure : dans l'exemple précédent, si l'on devait exprimer le temps d'attente considéré en minutes et non plus en heures, il faudrait adopter le paramètre
$$\lambda' = \frac{\lambda}{60} = \frac{12}{60} = 0,2$$
- Ce paramètre λ permettant de spécifier une loi exponentielle est de nature *intensive* et non plus *extensive* : dans notre exemple, il s'exprime en *(nb. de clients)/h*, et non en *nb. de clients*.

4.1.3 Variables normales

On dira que la variable continue $Z : \Omega \longrightarrow \mathbb{R}$ suit une ***loi normale standard*** (ou encore une ***loi gaussienne standard***) si sa fonction de densité f_Z est donnée par

$$\forall x \in \mathbb{R}, \quad f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Ici, *il s'avère impossible d'exprimer la fonction de répartition F_Z d'une telle variable comme composée de fonctions connues.*
Plutôt que de rechercher en vain une telle expression, on pourra s'appuyer sur les propriétés de symétrie de f_Z et F_Z :

$\forall x \in \mathbb{R}, \quad f_Z(-x) = f_Z(x)$ *(parité de f_Z)* et : $F_Z(-x) = 1 - F_Z(x)$ *(conséquence de la parité de f_Z)*

puis utiliser des valeurs approchées de $F_Z(x)$ pour x variant dans $\mathbb{R}_+$ (cf **Tables 2a et 2b** données en Appendice).

Pour des raisons qui seront bientôt abordées, les variables normales sont souvent utilisées pour modéliser des erreurs de mesure.
Exemple d'utilisation : Un chercheur en microbiologie doit mesurer le diamètre d'une colonie de bactéries dans un milieu donné. Il sait que la précision de ses instruments de mesure est loin d'être parfaite, et part du principe que l'erreur commise sur une telle mesure, *exprimée en dizièmes de mm*, suit une loi normale standard.
Il prend une telle mesure et trouve un valeur de 3mm pour ce diamètre. Quelle est la probabilité pour que la véritable valeur du diamètre soit située sous 2,9mm ?

$$\mathbb{P}\{Z \geq 1\} = 1 - \mathbb{P}\{Z < 1\} = 1 - F_Z(1) \approx 1 - 0,8413 \textit{(cf table !)}$$

cette probabilité vaut donc $15,87\%$.
Si l'on s'intéresse à la probabilité d'avoir une vraie valeur située entre 3,05mm et 3,25mm, il vient

$$\mathbb{P}\{-2,5 \leq Z \leq -0,5\} = \mathbb{P}\{Z \leq -0,5\} - \mathbb{P}\{Z < -2,5\} = F_X(2,5) - F_X(0,5) = \approx 0,9938 - 0,6915,$$

cette nouvelle probabilité vaut donc (approximativement) $30,23\%$.

Etant donné deux constantes $\mu \in \mathbb{R}$ et $\sigma > 0$, la variable

$$W = \sigma Z + \mu$$

obtenue à partir de la variable normale standard Z au moyen d'une dilatation (facteur positif σ) et d'une translation (constante additive μ) sera elle aussi considérée comme une variable normale. Sa loi se note $\mathcal{N}(\mu; \sigma^2)$. Il s'avère que la fonction de densité d'une telle variable W peut être exprimée explicitement :

$$\forall x \in \mathbb{R}, \quad f_W(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cependant, dans la pratique on se ramènera au cas standard afin de pouvoir utiliser à nouveau les tables de référence correspondant à la loi $\mathcal{N}(0; 1)$ (Tables 2a et 2b).

Exemple d'évaluation de probabilité normale :
Si W suit une loi $\mathcal{N}(\mu = 17; \sigma^2 = 144)$, la probabilité d'observer que $3 \leq W \leq 15$ vaut

$$\begin{aligned}
\mathbb{P}\{3 \leq W \leq 15\} &= \mathbb{P}\{(3-\mu) \leq (W-\mu) \leq (15-\mu)\} \\
&= \mathbb{P}\left\{\tfrac{(3-\mu)}{\sigma} \leq \tfrac{(W-\mu)}{\sigma} \leq \tfrac{(15-\mu)}{\sigma}\right\} \\
&= \mathbb{P}\left\{\tfrac{-14}{12} \leq Z \leq \tfrac{-2}{12}\right\} \\
&= F_Z(-\tfrac{1}{6}) - F_Z(-\tfrac{7}{6}) \\
&= F_Z(\tfrac{7}{6}) - F_Z(\tfrac{1}{6}) \\
&\approx 0,879 - 0,5675,
\end{aligned}$$

cette probabilité s'élève donc (approximativement) à $31,15\%$.

4.2 Calculs d'espérances et de variances

Le calcul de l'espérance d'une variable aléatoire continue X est entièrement analogue à celui de l'espérance d'une variable discrète D, en remplaçant les probabilités ponctuelles $\mathbb{P}\{D = x\}$ par des valeurs $f_X(x)$ de la fonction de densité, et la somme discrète $\sum_x x \cdot \mathbb{P}\{D = x\}$ par une somme continue, autrement dit par une intégrale :

$$\mathbb{E}[X] = \int_I x f_X(x) dx$$

Ici encore, $I = X(\Omega)$ est l'intervalle des valeurs envisageables pour X, et si I n'est pas borné, l'intégrale $\int_I x f_X(x)dx$ est définie par un passage à la limite. [1]
Tout comme pour une variable discrète, on pourra définir le moment d'ordre m de la variable continue X comme

$$\mathbb{E}[X^2] = \int_I x^2 f_X(x) dx,$$

et plus généralement son moment d'ordre m comme

$$\mathbb{E}[X^m] = \int_I x^m f_X(x) dx.$$

La variance de X n'est ensuite que la différence entre le deuxième moment de X et le carré de son premier moment :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_I x^2 f_X(x) dx - \left(\int_I x f_X(x) dx \right)^2$$

Tout comme dans le chapitre précédent, on peut affirmer que

1. L'espérance est *monotone* et *linéaire* :

$$(X \leq Y \implies \mathbb{E}[X] \leq \mathbb{E}[Y]) \text{ et } (\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y])$$

2. La variance est *quadratique*, *additive* ***dans le cas de variables indépendantes*** :

$$\big(\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)\big) \text{ et } (\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \text{ si } X \text{ et } Y \text{ sont indépendantes.})$$

Passons maintenant à quelques exemples.

[1] Il y a des fonctions de densité pour lesquelles une telle limite n'existe pas ou vaut $\pm\infty$; nous ne rencontrerons pas de telles fonctions, toutes les variables que nous étudierons posséderont des moments de tous ordres.

4.2.1 Variables uniformes

Pour une variable X distribuée uniformément sur l'intervalle $[a;b]$, on a

$$\mathbb{E}[X] = \int_a^b x f_X(x)dx = \int_a^b x \cdot \frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] = \frac{a+b}{2}$$

Pour le moment d'ordre 2 de X, on obtient ensuite

$$\mathbb{E}[X^2] = \int_a^b x^2 f_X(x)dx = \int_a^b x^2 \cdot \frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right] = \frac{a^2+ab+b^2}{3},$$

ce qui montre que

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}$$

4.2.2 Variables exponentielles

Si T est une variable exponentielle de paramètre λ, on a tout d'abord

$$\mathbb{E}[T] = \int_0^{+\infty} x f_T(x)dx = \int_0^{+\infty} x \cdot (\lambda e^{-\lambda x})dx = \lim_{A\to+\infty}\int_0^A x \cdot (\lambda e^{-\lambda x})dx$$

Observons qu'une intégration par parties donne ici

$$\begin{array}{rcl} \int_0^A x \cdot (\lambda e^{-\lambda x})dx & = & \left[x \cdot (-e^{-\lambda x})\right]_0^A - \int_0^A 1 \cdot (-e^{-\lambda x})dx \\ & = & \left[-Ae^{-\lambda A}\right] - \left[\frac{e^{-\lambda x}}{\lambda}\right]_0^A \\ & = & -Ae^{-\lambda A} - \frac{e^{-\lambda A}}{\lambda} + \frac{1}{\lambda} \end{array}$$

Ainsi :

$$\mathbb{E}[T] = \lim_{A\to+\infty}\int_0^A x \cdot (\lambda e^{-\lambda x})dx = \lim_{A\to+\infty}\left\{-Ae^{-\lambda A} - \frac{e^{-\lambda A}}{\lambda} + \frac{1}{\lambda}\right\} = \frac{1}{\lambda}$$

Le calcul du moment d'ordre 2 d'une telle variable peut aussi se faire au moyen d'une intégration par parties, on obtient

$$\mathbb{E}[T^2] = \int_0^{+\infty} x^2 f_T(x)dx = \lim_{A\to+\infty}\int_0^A x^2 \cdot (\lambda e^{-\lambda x})dx = \frac{2}{\lambda^2}$$

Il vient donc

$$\mathrm{Var}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{1}{\lambda^2}$$

4.2.3 Variables normales

Si Z est une variable normale standard, autrement une variable de loi $\mathcal{N}(0;1)$, on a tout d'abord

$$\mathbb{E}[Z] = \int_{-\infty}^{+\infty} x f_Z(x) dx = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \lim_{A \to +\infty} \int_{-A}^{A} x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0,$$

la fonction $x \mapsto \frac{x}{\sqrt{2\pi}} e^{-x^2/2}$ étant *impaire* !
L'évaluation du moment d'ordre 2 d'une telle variable nécessite un peu plus de travail et donne :

$$\mathbb{E}[Z^2] = \int_{-\infty}^{+\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 2 \int_{0}^{+\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$$

On obtient donc ensuite :

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = 1^2 - 0^2 = 1$$

Plus généralement, si W est une variable normale de loi $\mathcal{N}(\mu;\sigma^2)$, l'utilisation des propriétés de l'espérance et de la variance permet d'affirmer que

$$\mathbb{E}[W] = \mathbb{E}[\sigma Z + \mu] = \sigma \mathbb{E}[Z] + \mathbb{E}[\mu] = \sigma \cdot 0 + \mu = \mu$$

tandis que

$$\mathrm{Var}(W) = \mathrm{Var}(\sigma Z + \mu) = \mathrm{Var}(\sigma Z) = \sigma^2 \cdot \mathrm{Var}(Z) = \sigma^2$$

4.2.4 Tableau Récapitulatif

	$V(\Omega) =$	Loi de V	$\mathbb{E}[V]$	$\mathrm{Var}(V)$
$\mathcal{U}([a;b])$	$[a;b]$	$f_V(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{E}(\lambda)$	$]0;+\infty[$	$f_V(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathcal{N}(0;1)$	$\mathbb{R}$	$f_V(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	0	1
$\mathcal{N}(\mu;\sigma^2)$	$\mathbb{R}$	$f_V(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

4.3 Exercices d'Application

I Calculs avec des Variables Gaussiennes :

1. Lors d'un procès en attribution de paternité, un Professeur de Médecine indique que la durée (en jours) d'une grossesse suit approximativement une loi normale d'espérance $\mu = 270$ et de variance $\sigma^2 = 100$. L'un des pères putatifs parvient à prouver son absence du pays durant une période de cinquante jours s'étendant entre le 290ème et le 240ème jour précédant l'accouchement, et le tribunal décide ensuite de considérer qu'il ne *peut pas être* le père recherché.
   Quelle est la probabilité que le tribunal se soit trompé? (Donner une réponse précise s'appuyant sur le calcul de $\mathbb{P}\{240 < X < 290\}$, X désignant la variable aléatoire de durée de la grossesse).

2. Une usine produit des pièces en aluminium comportant une fente; la largeur de chaque fente produite suit une loi normale d'espérance $\mu = 2$ et d'écart-type $\sigma = 0,007$ (en cm), tandis que les limites de tolérance, permettant de considérer qu'une pièce n'est pas défectueuse, sont de

$$2,0000 \pm 0,0120 \text{ cm.}$$

Dans ces conditions, quel sera le pourcentage de pièces défectueuses?

3. On suppose que la taille, en centimètres, d'un homme âgé de 25 ans est une variable aléatoire normale d'espérance $\mu = 175$ et de variance $\sigma^2 = 36$. Quel est le pourcentage des hommes de 25 ans mesurant plus de 185 cm? Parmi les hommes de plus de 180 cm, quel est le pourcentage de ceux dépassant les 192 cm?

II Durées de travail :

On modélise la durée hebdomadaire de travail des cadres par une variable aléatoire de loi normale $N(m, \sigma^2)$ avec $m = 45h$.

1. Quelle est la proportion de cadres qui travaillent plus de $45h$ par semaine ?

2. Une étude a montré que 1% des cadres travaillent plus de $60h$ par semaine. En déduire la valeur de σ.

3. Compléter les affirmations suivantes :
- 5% des cadres travaillent moins de $...h$ par semaine ;
- 95% des cadres travaillent entre $...h$ et $...h$ par semaine (*plusieurs réponses possibles*).

III Utilisation de variables normales :

Un chef de service quitte son domicile à 8h45 et se rend, en voiture de sport, à son bureau qui ouvre à 9h. La durée du trajet est une variable $N(m, \sigma^2)$ avec $m = 12$ mn et $\sigma = 3$ mn. La secrétaire du service se rend elle aussi au bureau : elle prend le train à 7h45, puis marche 10mn pour prendre un autobus à 8h40 qui la dépose devant son bureau. Les durées des trajets en train et en autobus sont des variables normales $N(m_1, \sigma_1^2)$ et $N(m_2, \sigma_2^2)$ avec $m_1 = 42$ mn, $m_2 = 15$ mn, $\sigma_1 = 3$ mn, et $\sigma_2 = 4$ mn. Quelle est la probabilité

1. que le chef de service soit à l'heure?

2. que la secrétaire soit en retard?

3. que le chef de service soit à l'heure et la secrétaire en retard?

IV Encore une utilisation de variables normales :

Dans une usine de production automobile, on monte sur deux chaînes séparées les caisses de

voitures et les moteurs correspondants. On suppose que le montage d'une caisse commence à 6 heures et nécessite un temps aléatoire X (exprimé en minutes) distribué suivant une loi normale $N(60; 3^2)$. Le montage du moteur correspondant se fait sur une autre chaîne et nécessite un temps aléatoire Y (exprimé aussi en minutes) de loi $N(180; 4^2)$.

1. Avec quelle probabilité la caisse sera-t-elle terminée avant 6h55 ?

2. A quelle heure H doit-on commencer le montage du moteur si on veut qu'il soit terminé en moyenne au même instant que la caisse ?
   Dans la suite, on supposera que le montage du moteur commence effectivement à l'heure H.
   On utilisera aussi le théorème suivant: "Si Z_1 et Z_2 sont des variables aléatoires indépendantes et de lois normales, et si a et b sont des nombres réels, alors $aZ_1 + bZ_2$ suit encore une loi normale".

3. Combien valent l'espérance et la variance de $Y - X$? Quelle est la loi de $Y - X$?

4. Quelle est la probabilité que la caisse soit prête avant le moteur ? que la caisse et le moteur soit prêts en même temps ?

5. Les deux chaînes se rejoignent en une troisième qui doit terminer l'assemblage. Cet assemblage ne peut se faire que si la caisse et le moteur se présentent à des instants distants de moins de 5 minutes. Avec quelle probabilité l'assemblage aura-t-il lieu ?

6. Quel écart maximum faudrait-il tolérer entre les instants d'arrivée des deux parties sur la troisième chaîne pour que l'assemblage puisse être réalisé dans 90% des cas ?

V Voyages, voyages :

La durée prévue pour le trajet Paris-Orléans en train est $d = 60$ minutes. Dans les faits, cette durée, exprimée en minutes, est une variable aléatoire D qui vaut d dans 90% des cas et vaut $d + |X|$ dans les autres cas, X étant une variable aléatoire de loi normale $N(0, \sigma^2)$ avec $\sigma = 15$.

1. A partir de la table de loi $N(0, 1)$ et en vous aidant éventuellement d'un schéma,
   a- calculer $\mathbb{P}(X \leq 15)$ et $\mathbb{P}(X > 15)$
   b-justifier que $\mathbb{P}(|X| > 15) = 0.32$; $\mathbb{P}(|X| > 30) = 0.05$; $\mathbb{P}(|X| > 45) = 0.003$
   1) a- Quelle est la probabilité qu'un train, parti à l'heure de Paris, soit en retard à l'arrivée à Orléans?
   b- Pour un train en retard, quelle est la probabilité que le retard soit supérieur à 15 minutes?
   c- Etablir (en le justifiant) que la probabilité qu'un train, parti à l'heure de Paris,

accuse à l'arrivée à Orléans un retard de plus de 15 minutes est égale à $p = 0.032$.

2. Si le retard est supérieur à 15 minutes, la SNCF rembourse à chaque passager une somme forfaitaire de 10 euros. On note Y le montant du remboursement à prévoir pour chaque voyageur transporté.
   a- Quelle est la loi de Y? (justifier la réponse)
   b- Calculer le coût moyen du remboursement que la SNCF doit prévoir pour chaque passager.
   c- Quelle devrait être la valeur de σ pour que ce coût moyen soit divisé par 10 ?

3. Deux trains partent en même temps de Paris sur deux voies parallèles.
   a- Quelle est la probabilité que le train le plus rapide effectue le parcours en 60 minutes?
   b- Quelle est la probabilité que le train le moins rapide arrive avec un retard de plus de 15 minutes?

VI Calculs avec des Variables Exponentielles :

1. La durée de vie, exprimée en heures, d'un composant électronique définit une variable aléatoire continue V qui suit une loi exponentielle. On a constaté expérimentalement que 97,04% de ces composants fonctionnent encore au bout de 30 000 heures.
   - Montrer que cette constatation permet de fixer à 1.10^{-6} le paramètre de cette loi exponentielle.
   - Quelle est la probabilité qu'un composant fonctionne encore au bout de 60 000 heures? qu'il tombe en panne avant sa 10 000ème heure?
   - On suppose dans cette question que 10 composants sont montés en série : le système ne fonctionne que si les 10 composants fonctionnent.
     a- Calculer la probabilité que le système fonctionne encore au bout de 10^6 heures.
     b- Calculer la fonction de répartition de la variable aléatoire égale à la durée de vie du système; en déduire sa loi.
     c- Compléter la phrase: "Le système dure en moyenne ... fois ... (moins/plus) longtemps que chaque composant pris indépendamment".
   - On suppose maintenant que le système est constitué de 2 composants montés en redondance : le système fonctionne tant que l'un au moins des composants fonctionne. Reprendre la question b- ci-dessus, et calculer la 1/2 vie du système, c'est-à-dire la médiane de la durée de vie du système.

2. On modélise le temps écoulé entre les instants de passage de deux autobus d'une même ligne à un arrêt donné par une variable aléatoire X. Ce temps est mesuré en minutes et on suppose que X suit une loi exponentielle de paramètre $\lambda = 0,1$.
   - Calculer le temps moyen entre les instants de passage de deux bus.

- Alice arrive à l'arrêt juste après le passage d'un bus. Calculer la probabilité pour qu'elle attende plus de 10 minutes le bus suivant.
- Bob arrive 5 minutes après Alice à l'arrêt. Calculer la probabilité pour qu'Alice soit toujours là et qu'elle attende moins de 5 minutes le prochain bus.
- Bob arrive 5 minutes après Alice à l'arrêt. Sachant qu'Alice est toujours là, quelle est la probabilité que Bob attende plus de 10 minutes?

VII Calculs avec d'autres types de Variables :

1. On suppose que la v.a. X modélisant l'âge d'un PC arrivant au service après-vente d'un certain constructeur suit une loi continue de densité f_X telle que

$$f_X(x) = \begin{cases} k(x - x^3) & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

(La garantie des PC considérés dure un an, X est donc à valeurs dans $[0; 1]$).

- Montrer que $k = 4$.
- Etablir que X a pour espérance $\mathbb{E}(X) = \frac{8}{15}$ et pour variance $\text{Var}(X) = \frac{11}{225}$.
- Montrer que la fonction de répartition F_X de cette variable X est telle que

$$\forall x \in [0; 1], \quad F_X(x) = 2x^2 - x^4$$

- Rechercher les valeurs de $\mathbb{P}(X < 0,5)$ et de $\mathbb{P}(X < 0,6)$.

2. Soient X, Y, Z des variables aléatoires continues ayant pour densités les fonctions f_X, f_Y, f_Z données par

$$f_X(x) = \begin{cases} a & \text{si } -1 \leq x \leq 1 \\ 0 & \text{sinon,} \end{cases} \quad f_Y(x) = \begin{cases} b(1 - x^2) & \text{si } -1 \leq x \leq 1 \\ 0 & \text{sinon} \end{cases}$$

et

$$f_Z(x) = \begin{cases} c(1 + x) & \text{si } -1 \leq x < 0 \\ c(1 - x) & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

- Rechercher les valeurs précises des réels a, b, c, puis dessiner dans un même repère les courbes représentatives de ces trois fonctions de densité (on pourra préciser, pour chaque courbe, quels sont les points d'abscisses -1, $-0,5$, 0, $0,5$, 1).
- Utiliser ces trois courbes pour classer par ordre croissant les trois probabilités

$$\mathbb{P}\{-0,5 \leq X \leq 0,5\}, \mathbb{P}\{-0,5 \leq Y \leq 0,5\} \text{ et } \mathbb{P}\{-0,5 \leq Z \leq 0,5\}.$$

ou encore les trois probabilités

$$\mathbb{P}\{|X| > 0,5\}, \mathbb{P}\{|Y| > 0,5\} \text{ et } \mathbb{P}\{|Z| > 0,5\}.$$

- Retrouver les comparaisons de la question précédente en calculant précisément ces probabilités.
- Trouver les valeurs de $\mathbb{E}(X), \mathbb{E}(Y), \mathbb{E}(Z)$ sans aucun calcul.
- Ranger par ordre croissant les variances $\mathrm{Var}(X), \mathrm{Var}(Y)$ et $\mathrm{Var}(Z)$ en comparant les graphes des densités f_X, f_Y et f_Z.
- Retrouver les comparaisons de la question précédente en calculant précisément ces variances.

VIII Loi uniforme dans $[0,1]$ ⊞ :

Dans cet exercice on appellera "nombre au hasard dans $[0,1]$" un nombre obtenu comme la réalisation d'une variable aléatoire de loi uniforme dans $[0,1]$. La fonction `rand` de `Scilab` répond à cette demande (*voir l'aide pour la syntaxe*).

1. Simulez une colonne de 10 nombres au hasard dans $[0,1]$ et tracez un histogramme de 10 intervalles représentant la répartition de ces 10 valeurs.
   Recommencez avec 100 nombres, puis avec 1000. Que constatez-vous?
2. Pour $n = 1$ jusqu'à $n = 1000$, calculez la moyenne des n premiers nombres obtenus à la question précédente, ainsi que leur variance.
   Tracez en fonction de n, les moyennes et les variances obtenues. Que constatez-vous?
   (*L'espérance, c'est à dire la moyenne théorique, d'une v.a. de loi uniforme dans* $[0,1]$ *vaut 0.5 et la variance théorique vaut* $\frac{1}{12} = 0.833...$)
3. Simulez 1000 échantillons de taille 100 de nombres au hasard dans $[0,1]$ (*il est inutile et même déconseillé d'afficher le résultat!*).
   Calculez la moyenne de chaque échantillon et tracez l'histogramme des moyennes ainsi calculées.
   Que constatez-vous ?

IX Simulations de lois uniformes ⊞ :

1. Simuler un échantillon de taille 1000 de nombres au hasard de loi uniforme dans $[-1,1]$.
   Calculer la moyenne de cet échantillon et tracer un histogramme constitué de 10 intervalles de même largeur.
2. En utilisant la fonction `rand` composée avec une fonction qui, par dilatation et translation, transforme l'intervalle $[0,1]$ en $[-1,1]$, simuler un échantillon de taille 1000 de nombres au hasard dans $[-1,1]$.
   Calculer la moyenne et tracer un histogramme constitué de 10 intervalles de même largeur.
   Cet échantillon a-t-il la même distribution que celui de la question a)?
3. Simuler deux échantillons de taille 1000 de loi uniforme dans $[-1,1]$ (méthode au choix) et effectuer les moyennes terme à terme des 2 échantillons. On obtient ainsi un nouvel échantillon de taille 1000.

Entre quelles bornes se trouvent les valeurs de cet échantillon? Quelle est sa valeur moyenne?
Tracer l'histogramme. Le comparer aux ceux des questions précédentes. Ce 3ème échantillon a-t-il la même distribution que les deux premiers?

X Calculs avec une Variable Normale Standard :
Soit Z une variable aléatoire suivant une loi $\mathcal{N}(0;1)$.
QCM1 : En acceptant une marge d'erreur de $\pm 0,01$, quelles sont les affirmations exactes parmi celles données ci-dessous ?

A $\mathbb{P}(Z < 1) = 0,84$
B $\mathbb{P}(Z > 1,5) = 0,93$
C $\mathbb{P}(Z < -2) = 0,05$
D $\mathbb{P}(Z > -0,5) = 0,31$
E $\mathbb{P}(Z > -0,5) = 0,69$

QCM2 : Toujours en acceptant une marge d'erreur de $\pm 0,01$, quelles sont les affirmations exactes parmi celles données ci-dessous ?

A $\mathbb{P}(1 < Z < 1,5) = 0,09$
B $\mathbb{P}(-1 < Z < 1,5) = 0,77$
C $\mathbb{P}(-2 < Z < -1) = 0,27$
D $\mathbb{P}(|Z| > 1) = 0,32$
E $\mathbb{P}(|Z| > 0,5) = 0,68$

QCM3 : Toujours à $0,01$ près, quelles sont les affirmations exactes parmi celles données ci-dessous ?

A Si $\mathbb{P}(Z < z) = 0,8$, on a $z = 1,28$
B Si $\mathbb{P}(Z > z) = 0,1$, on a $z = 0,84$
C Si $\mathbb{P}(Z > z) = 0,95$, on a $z = -1,65$
D Si $\mathbb{P}(Z < z) = 0,35$, on a $z = -0,39$
E Si $\mathbb{P}(|Z| > z) = 0,6$, on a $z = 1,96$

V Calculs avec une Variable Normale Non-Standard :
Dans une certaine population, la concentration d'une substance S suit une loi normale de moyenne $\mu = 100$ et d'écart-type $\sigma = 10$.
QCM4 : Quelles sont les affirmations exactes parmi celles données ci-dessous, à 1% près ?

A Dans cette population, la proportion d'individus qui ont une concentration de la substance S supérieure à 110 est de 46%.

B Dans cette population, la proportion d'individus qui ont une concentration de la substance S supérieure à 110 est de 16%.

C Dans cette population, la proportion d'individus qui ont une concentration de la substance S inférieure à 85 est de 16%.

D Dans cette population, la proportion d'individus qui ont une concentration de la substance S inférieure à 85 est de 7%.

E Dans cette population, la proportion d'individus qui ont une concentration de la substance S comprise entre 85 et 110 est de 38%.

QCM5 : On souhaite définir deux seuils pour la concentration de cette substance chez un individu : celui-ci sera déclaré *"malade"* si la concentration dépasse le seuil s_1, ou encore comme faisant partie d'une sous-population *"à risque"* si cette concentration dépasse un autre seuil s_2. L'objectif est de définir ces seuils s_1 et s_2 en sorte que 5% des individus soient déclarés malades et 5% des individus déclarés comme faisant partie de la sous-population à risque.
Peut-on alors affirmer que

A $s_1 = 116,45$?

B $s_1 = 119,60$?

C $s_2 = 106,74$?

D $s_2 = 108,42$?

E $s_2 = 112,82$?

XI Calculs avec des Variables Normales Non-Standard :

Une maladie se présente sous deux formes cliniques, une forme modérée et une forme sévère. Dans la forme modérée, la concentration d'une certaine substance suit une loi normale de moyenne $\mu = 90$ et d'écart-type $\sigma = 10$. Dans la forme sévère, la concentration de cette même substance suit une loi normale de moyenne $\mu' = 105$ et d'écart-type $\sigma' = 15$. On note X la concentration de cette substance chez un malade, mais aussi M et S les événements *"Etre atteint de la forme modérée"* et *"Etre atteint de la forme sévère"*, respectivement.
QCM6 : Quelles sont les affirmations exactes parmi celles données ci-dessous, à $0,1\%$ près ?

A $\mathbb{P}(X < 95|M) = 0,309$

B $\mathbb{P}(X < 95|M) = 0,691$

C $\mathbb{P}(X < 95|S) = 0,251$

D $\mathbb{P}(X < 95|S) = 0,749$

E $\mathbb{P}(X < 95|S) = 0,841$

QCM7 : Parmi les patients atteints de cette maladie, on constate que 70% ont la forme modérée et 30% la forme sévère. Peut-on alors affirmer que

A $\mathbb{P}(X > 95) = 0,441$?

B $\mathbb{P}(X > 95) = 0,559$?

C $\mathbb{P}(M|X > 95) = 0,166$?

D $\mathbb{P}(S|X > 95) = 0,510$?

E $\mathbb{P}(S|X > 95) = 0,834$?

XII Calculs avec des Variables de Durée de Vie Exponentielles :

On considère des ampoules ayant une durée de vie médiane de 100 heures, ce qui signifie que la probabilité qu'une telle ampoule soit toujours en état de fonctionner après 100 heures d'utilisation vaut précisément $0,5$. On suppose que la durée de vie T d'une telle ampoule suit une loi exponentielle de paramètre $a > 0$, et l'on note F la fonction de répartition de T et S sa fonction de survie :

$$\forall t \in \mathbb{R}_+, \quad F(t) = \mathbb{P}(T \leq t),\ S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t).$$

QCM8 : Quelles sont les affirmations exactes parmi celles données ci-dessous ?

A $S(100) = 0,5$

B $S(100) = F(100)$

C $a = 0,007$

D $a = 0,003$

E $\mathbb{E}(T) = 333$

QCM9 : Soit p la probabilité qu'une telle ampoule grille durant sa première heure de fonctionnement. En utilisant l'approximation $e^x \approx 1 + x$, valable pour $x \approx 0$, peut-on affirmer que

A $p = 0,001$?

B $p = 0,003$?

C $p = 0,005$?

D $p = 0,007$?

E $p = 0,009$?

QCM10 : On allume une centaine d'ampoules simultanément. Soit N la variable comptant le nombre d'ampoules ayant grillé durant la première heure de fonctionnement. Peut-on affirmer que

A N suit une loi binomiale ?

B N suit approximativement une loi de Poisson ?

C $\mathbb{E}(N) = 7$?

D $\mathbb{P}(N = 0) = 0,35$ à $0,01$ près ?

E $\mathbb{P}(N = 0) = 0,50$ à $0,01$ près ?

XIII Calculs avec une Combinaison de Variables Normales :

On étudie une maladie survenant seulement si un certain gène subit une mutation. Cette mutation peut se produire sur un seul allèle (cas d'un malade *hétérozygote*) ou sur les deux allèles (cas d'un malade *homozygote*).

On s'intéresse à une population d'individus ayant un ou deux allèles mutés ; l'âge de la maladie est une variable aléatoire suivant une loi qui dépend du nombre d'allèles mutés :

-pour les individus homozygotes, l'âge de début de maladie exprimé en années suit une loi normale de moyenne $\mu = 25$ et décart-type $\sigma = 10$.

-pour les individus hétérozygotes, l'âge de début de maladie exprimé en années suit une loi normale de moyenne $\mu' = 40$ et décart-type $\sigma' = 15$.

La proportion d'individus homozygotes dans cette population est de 20%, celle des individus hétérozygotes est de 80%.

QCM11 : En appelant f la fonction de densité de cette variable d'âge de la maladie, g la fonction de densité de cette variable d'âge au sein de la population des homozygotes et h la fonction de densité de cette variable d'âge au sein de la population des hétérozygotes, on a

$$\forall x \in \mathbb{R}, \quad f(x) = p \cdot g(x) + (1-p) \cdot h(x), \quad \text{où}$$

A $g(x)$ est une densité de loi normale

B $h(x)$ est une densité de loi normale

C $f(x)$ est une densité de loi normale

D $p = 0,20$

E $p = 0,50$

QCM12 : Peut-on affirmer que

A $\mathbb{E}(X) = 40$ ans chez les homozygotes

B $\mathbb{E}(X) = 40$ ans chez les hétérozygotes

C $\mathbb{E}(X) = 30$ ans chez les individus ayant au moins une mutation

D $\mathbb{E}(X) = 32,5$ ans chez les individus ayant au moins une mutation

E $\mathbb{E}(X) = 37$ ans chez les individus ayant au moins une mutation

QCM13 : Peut-on affirmer que

A $\mathbb{E}(X^2) = 1369$ ans^2 chez les individus ayant au moins une mutation

B $\mathbb{E}(X^2) = 1605$ ans^2 chez les individus ayant au moins une mutation

C $\mathrm{Var}(X) = 200$ ans^2 chez les individus ayant au moins une mutation

D $\mathrm{Var}(X) = 236$ ans^2 chez les individus ayant au moins une mutation

E $\mathrm{Var}(X) = 549$ ans^2 chez les individus ayant au moins une mutation

QCM14 : A $0,01$ près, a-t-on

A $\mathbb{P}(X < 30) = 0,69$ chez les homozygotes ?

B $\mathbb{P}(X < 30) = 0,75$ chez les homozygotes ?

C $\mathbb{P}(X < 30) = 0,25$ chez les hétérozygotes ?

D $\mathbb{P}(X < 30) = 0,31$ chez les hétérozygotes ?

E $\mathbb{P}(X < 30) = 0,39$ chez les hétérozygotes ?

QCM15 : A $0,01$ près, peut-on affirmer que

A $\mathbb{P}(X < 30) = 0,34$ chez les individus ayant au moins une mutation ?

B $\mathbb{P}(X < 30) = 0,40$ chez les individus ayant au moins une mutation ?

C $\mathbb{P}(X < 30) = 0,46$ chez les individus ayant au moins une mutation ?

D Si un individu est atteint de la maladie avant d'avoir 30 ans, la probabilité qu'il soit homozygote vaut $0,38$?

E Si un individu est atteint de la maladie avant d'avoir 30 ans, la probabilité qu'il soit homozygote vaut $0,41$?

XIV Calculs avec des Variables Normales Non-Standard :

Pour un patient atteint de la maladie $\mathcal{M}$, la concentration plasmatique X d'une certaine molécule suit une loi normale de moyenne $\mu = 25$ et d'écart-type $\sigma = 7$.

QCM16 : A $0,01$ près, quelles sont les affirmations exactes parmi celles données ci-dessous ?

A $\mathbb{P}(X > 25) = 0,5$

B $\mathbb{P}(X < 20) = 0,24$

C $\mathbb{P}(X > 32) = 0,16$

D $\mathbb{P}(18 < X < 32) = 0,58$

E $\mathbb{P}(18 < X < 32) = 0,68$

QCM17 : On suppose que dans la population des individus qui ne sont pas atteints de la maladie $\mathcal{M}$, cette concentration plasmatique X suit encore une loi normale, mais avec une moyenne $\mu' = 28$ et d'écart-type $\sigma' = 6$.
L'événement "L'individu a la maladie $\mathcal{M}$" est noté M, et l'on suppose que $\mathbb{P}(M) = 0,5$.
A $0,01$ près, peut-on affirmer que

A $\mathbb{P}(X > 30|M) = 0,45$?

B $\mathbb{P}(X > 30|\overline{M}) = 0,37$?

C $\mathbb{P}(M|X > 30) = 0,23$?

D $\mathbb{P}(M|X > 30) = 0,39$?

E $\mathbb{P}(M|X > 30) = 0,58$?

QCM18 : On mesure la valeur de X indirectement à l'aide d'un appareil fournissant une valeur Y telle que $Y = 3,5\,X^2$.
Parmi les intervalles suivants, lesquels contiennent avec probabilité $0,8$ la valeur de Y pour une mesure effectuée sur un individu malade ?

A $]-\infty; 3340]$

B $]-\infty; 4040]$

C $[-4040; 4040]$

D $[-4667; 3340]$

E $[-4667; 3641]$

XV Calculs avec une Variable Uniforme :

Soit U une variable aléatoire continue distribuée uniformément sur l'intervalle $[1; 3]$. Alors : **QCM19 :** La probabilité que X soit solution de l'équation

$$x^2 - 3x + 2 = 0$$

vaut

A 0

B $0,25$

C $0,5$

D $0,75$

E 1

QCM20 : La probabilité que X soit solution de l'inéquation

$$x^2 - 3x + 2 < 0$$

vaut

A 0

B $0,25$

C $0,5$

D $0,75$

E 1

5 STATISTIQUE INFERENTIELLE (I) : TCL et INTERVALLES DE CONFIANCE

Dans cet avant- dernier chapitre, on adopte à nouveau le point de vue de la Statistique en se situant *juste après* la réalisation d'un certain nombre d'expériences aléatoires dont les résultats ont été enregistrés.

5.1 Méthode du Maximum de Vraisemblance

Considérons une suite $X_1, X_2, \ldots, X_n$ de variables aléatoires ***indépendantes et de même loi***. [2] La nature de la loi commune à ces variables (binomiale, Poisson, uniforme ...) nous est connue, mais pas le ou les paramètre(s) précis de cette loi (n et p dans le contexte binomial, λ dans le contexte Poissonien, a et b pour une variable uniforme ...). Les expériences aléatoires correspondant à ces variables viennent d'être réalisées, en sorte que l'on a pu enregistrer les observations

$$X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n$$

Dans ces conditions, quelle est la valeur la plus vraisemblable du ou des paramètre(s) inconnu(s) ?

5.1.1 Cas d'un échantillon discret

Supposons que les variables i.i.d. $X_1, X_2, \ldots, X_n$ sont de nature discrète, et que leur loi commune dépend d'un certain paramètre θ.
Ayant enregistré les valeurs $x_1, x_2, \ldots, x_n$ prises par ces variables, on est en mesure de définir la ***fonction de vraisemblance*** V par

$$V_{x_1,x_2,\ldots,x_n}(\theta) = \mathbb{P}\{X_1 = x_1\} \times \mathbb{P}\{X_2 = x_2\} \times \ldots \times \mathbb{P}\{X_n = x_n\}$$

Il est à noter que cette fonction V dépend du nombre θ et non des valeurs $x_1, x_2, \ldots, x_n$, qui ont été enregistrées une fois pour toutes. Cette dépendance en θ se fait sentir dans l'évaluation de chacune des probabilités $\mathbb{P}\{X_k = x_k\}$, que l'on pourra donc aussi écrire $\mathbb{P}_\theta\{X_k = x_k\}$.
La valeur la plus vraisemblable du paramètre θ est alors celle qui maximise $V_{x_1,x_2,\ldots,x_n}(\theta)$.

1er Exemple : estimation de p pour un échantillon binomial
Si les variables i.i.d. $X_1, X_2, \ldots, X_n$ sont des variables binomiales $\mathcal{B}(N; p)$, où N est connu et p inconnu, comment estimer la valeur du paramètre p au vu des observations $X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n$?

[2] On parlera aussi d'échantillon ***i.i.d.***, l'acronyme i.i.d. signifiant ***indépendates et identiquement distribuées***.

La fonction de vraisemblance vaut ici

$$\begin{aligned} V_{x_1,x_2,\ldots,x_n}(p) &= \mathbb{P}\{X_1 = x_1\} \times \mathbb{P}\{X_2 = x_2\} \times \ldots \times \mathbb{P}\{X_n = x_n\} \\ &= \binom{N}{x_1} p^{x_1}(1-p)^{N-x_1} \times \binom{N}{x_2} p^{x_2}(1-p)^{N-x_2} \times \ldots \times \binom{N}{x_n} p^{x_n}(1-p)^{N-x_n} \\ &= \left(\binom{N}{x_1}\binom{N}{x_2}\ldots\binom{N}{x_n}\right) \times p^{x_1+x_2+\ldots+x_n} \times (1-p)^{nN-(x_1+x_2+\ldots+x_n)} \end{aligned}$$

p étant la seule variable dans cette dernière expression, le produit $\left(\binom{N}{x_1}\binom{N}{x_2}\ldots\binom{N}{x_n}\right)$ peut être considéré comme un facteur constant K, il vient donc

$$V_{x_1,x_2,\ldots,x_n}(p) = K \times p^{\sum_{k=1}^n x_k} \times (1-p)^{nN-\sum_{k=1}^n x_k}$$

Puisque la fonction ln est croissante, maximiser la vraisemblance V revient à maximiser son logarithme ; on a

$$\ln\left[V_{x_1,x_2,\ldots,x_n}(p)\right] = \ln K + \left(\sum_{k=1}^n x_k\right)\ln p + \left(nN - \sum_{k=1}^n x_k\right)\ln(1-p)$$

Il reste à dériver cette dernière expression en p pour voir que la valeur optimale de p satisfait

$$\frac{\sum_{k=1}^n x_k}{p} = \frac{nN - \sum_{k=1}^n x_k}{1-p}$$

L'unique valeur de p permettant de maximiser la vraisemblance est alors tout simplement

$$\hat{p} = \frac{1}{nN}\sum_{k=1}^n x_k = \frac{1}{N}\left(\frac{1}{n}\sum_{k=1}^n x_k\right) = \frac{1}{N}\overline{x}_n$$

2ème Exemple : estimation de λ pour un échantillon Poissonien

Si maintenant les variables i.i.d. $X_1, X_2, \ldots, X_n$ sont poissoniennes de paramètre λ, où λ est inconnu, comment estimer λ au vu des observations $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$?
La fonction de vraisemblance est maintenant telle que

$$\begin{aligned} V_{x_1,x_2,\ldots,x_n}(\lambda) &= \mathbb{P}\{X_1 = x_1\} \times \mathbb{P}\{X_2 = x_2\} \times \ldots \times \mathbb{P}\{X_n = x_n\} \\ &= \left(e^{-\lambda}(\tfrac{\lambda^{x_1}}{x_1!})\right) \times \left(e^{-\lambda}(\tfrac{\lambda^{x_2}}{x_2!})\right) \times \ldots \times \left(e^{-\lambda}(\tfrac{\lambda^{x_n}}{x_n!})\right) \\ &= \tfrac{1}{x_1!x_2!\ldots x_n!} \times e^{-n\lambda} \times \lambda^{\sum_{k=1}^n x_k} \end{aligned}$$

Dans cette dernière expression, $\left(\frac{1}{x_1!x_2!\ldots x_n!}\right)$ joue le rôle d'un facteur constant K.
On a ensuite :

$$\ln\left[V_{x_1,x_2,\ldots,x_n}(\lambda)\right] = \ln K - n\lambda + (\sum_{k=1}^n x_k)\ln\lambda$$

En dérivant cette dernière expression, on voit que la valeur optimale $\hat{\lambda}$ du paramètre λ satisfait

$$-n + \frac{\sum_{k=1}^n x_k}{\hat{\lambda}} = 0$$

L'estimation $\hat{\lambda}$ du paramètre λ par la méthode du maximum de vraisemblance nous donne donc tout simplement $\hat{\lambda} = \hat{\lambda}(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{k=1}^{n} x_k = \overline{x}_n$ dans ce contexte poissonien.

3ème Exemple : estimation des paramètres a et b pour un échantillon uniforme
Supposons à présent que les variables i.i.d. $X_1, X_2, \ldots, X_n$ sont uniformes sur un intervalle discret $[\![a; b]\!]$ dont les extrémités a et b nous sont inconnues. Au vu des observations $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$, quel est le couple $(a; b)$ le plus vraisemblable ?
Remarquons que la vraisemblance d'un couple $(a; b)$ est nulle dès lors que $a > X_i$ pour un certain indice i, ou encore dès lors que $b < X_j$ pour un certain indice j. On a plus précisément

$$\begin{aligned} V_{x_1,x_2,\ldots,x_n}(a;b) &= \mathbb{P}\{X_1 = x_1\} \times \mathbb{P}\{X_2 = x_2\} \times \ldots \times \mathbb{P}\{X_n = x_n\} \\ &= \begin{cases} 0 & \text{si } a > \min(x_1; x_2; \ldots; x_n) \text{ ou } b < \max(x_1; x_2; \ldots; x_n) \\ \left(\frac{1}{b-a}\right)^n & \text{sinon} \end{cases} \end{aligned}$$

Pour maximiser cette vraisemblance V, il faut donc faire en sorte que la longueur $(b - a)$ de l'intervalle discret $[\![a; b]\!]$ soit aussi petite que possible, tout en veillant à ce que la valeur de V ne devienne pas nulle ($a \leq \min(x_1; x_2; \ldots; x_n)$ et $b \geq \max(x_1; x_2; \ldots; x_n)$). Il vient donc ici

$$\hat{a} = \hat{a}(x_1; x_2; \ldots; x_n) = \min(x_1; x_2; \ldots; x_n) \text{ et } \hat{b} = \hat{b}(x_1; x_2; \ldots; x_n) = \max(x_1; x_2; \ldots; x_n)$$

5.1.2 Cas d'un échantillon continu

Supposons à présent que les variables i.i.d. $X_1, X_2, \ldots, X_n$ sont de nature continue et que leur loi commune dépend d'un certain paramètre θ.
Ayant enregistré les valeurs $x_1, x_2, \ldots, x_n$ prises par ces variables, il convient de définir la ***fonction de vraisemblance*** V comme produit de "probabilités infinitésimales", autrement dit comme produit de densités :

$$V_{x_1,x_2,\ldots,x_n}(\theta) = f_{X_1}(x_1) \times f_{X_2}(x_2) \times \ldots \times f_{X_n}(x_n)$$

Notons que l'on fait usage d'une seule et même fonction de densité dans ce dernier produit, puisque les variables $X_1, X_2, \ldots, X_n$ sont identiquement distribuées, et que l'expression de cette fonction de densité fait apparaître le paramètre inconnu θ. On pourrait donc tout aussi bien écrire

$$V_{x_1,x_2,\ldots,x_n}(\theta) = f_\theta(x_1) \times f_\theta(x_2) \times \ldots \times f_\theta(x_n),$$

et la valeur estimée $\hat{\theta}$ du paramètre inconnu est à nouveau celle qui maximise $V_{x_1,x_2,\ldots,x_n}(\theta)$.

1er Exemple : estimation de μ pour un échantillon normal
Soit $X_1, X_2, \ldots, X_n$ un échantillon i.i.d. de variables normales. On suppose que l'espérance μ commune à ces variables nous est inconnue, tandis que leur variance commune σ^2 est connue et vaut 1. Comment estimer la véritable valeur de μ à partir des observations $X_1 = x_1, X_2 =$

$x_2, \ldots, X_n = x_n$?
La vraisemblance est ici définie par

$$\begin{aligned} V_{x_1,x_2,\ldots,x_n}(\mu) &= f_\mu(x_1) \times f_\mu(x_2) \times \ldots \times f_\mu(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2}}\right) \times \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2}}\right) \times \ldots \times \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_n-\mu)^2}{2}}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \times e^{-\frac{1}{2}\sum_{k=1}^n (x_k-\mu)^2} \end{aligned}$$

En posant $K = \left(\frac{1}{\sqrt{2\pi}}\right)^n$ et en passant au logarithme de la vraisemblance, il vient

$$\ln\left(V_{x_1,x_2,\ldots,x_n}(\mu)\right) = \ln K - \frac{1}{2}\sum_{k=1}^n (x_k - \mu)^2$$

Une dérivation de cette dernière expression par rapport à μ montre ensuite que la vraisemblance est maximale lorsque

$$\sum_{k=1}^n (x_k - \mu) = 0,$$

c'est à dire pour

$$\mu = \hat{\mu}(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{k=1}^n x_k = \overline{x}_n$$

L'estimation du paramètre μ se fait donc à nouveau par la ***moyenne arithmétique*** $\overline{x}_n = \frac{1}{n}\sum_{k=1}^n x_k$ des résultats observés, que l'on appelle aussi ***moyenne empirique***.

2ème Exemple : estimation de λ pour un échantillon exponentiel
Considérons enfin la situation où $X_1, X_2, \ldots, X_n$ est un échantillon i.i.d. de variables exponentielles dont le paramètre λ nous est inconnu. Comment estimer λ par la méthode du maximum de vraisemblance, en s'appuyant sur les observations $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$?
La vraisemblance est maintenant définie par

$$\begin{aligned} V_{x_1,x_2,\ldots,x_n}(\lambda) &= f_\lambda(x_1) \times f_\lambda(x_2) \times \ldots \times f_\lambda(x_n) \\ &= \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda x_2} \times \ldots \times \lambda e^{-\lambda x_n} \\ &= \lambda^n \times e^{-\lambda \sum_{k=1}^n x_k}, \end{aligned}$$

son logarithme est donc donné par

$$\ln V_{x_1,x_2,\ldots,x_n}(\lambda) = n \ln \lambda - \left(\sum_{k=1}^n x_k\right)\lambda$$

Il reste à constater que la dérivée en λ de ce logarithme s'annule seulement lorsque

$$\frac{n}{\lambda} = \sum_{k=1}^n x_k$$

On obtient donc en fin de compte

$$\hat{\lambda} = \hat{\lambda}(x_1, x_2, \ldots, x_n) = \frac{n}{\sum_{k=1}^{n} x_k}$$

comme estimation de λ par la méthode du maximum de vraisemblance.

5.1.3 Biais d'un estimateur

Supposons que l'on cherche à nouveau à estimer la véritable valeur du paramètre θ apparaissant dans une loi commune à chacune des variables $X_1, X_2, \ldots, X_n$ d'un échantillon i.i.d., et que l'on utilise pour cela une certaine fonction

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$$

des observations $x_1, x_2, \ldots, x_n$ enregistrées.
Le ***biais*** de cet estimateur ponctuel $\hat{\theta}$ est alors défini comme la différence entre l'espérance mathématique de $\hat{\theta}(X_1, X_2, \ldots, X_n)$ et la valeur-cible θ :

$$\text{Biais}_{\hat{\theta}}(\theta) = \mathbb{E}_\theta\left[\hat{\theta}(X_1, X_2, \ldots, X_n)\right] - \theta$$

Le Biais peut donc être considéré comme une fonction de θ.
Lorsque cette fonction est identiquement nulle, on parlera d'***estimateur sans biais***, dans le cas contraire d'***estimateur biaisé***.
Voyons tout de suite que les estimateurs obtenus par la méthode du max. de vraisemblance *ne sont pas toujours sans biais* !

1. Si $X_1, X_2, \ldots, X_n$ est un échantillon i.i.d. de variables uniformes sur l'intervalle $[0; b]$, on vérifie, comme on l'a fait dans le cadre discret, que l'estimateur $\hat{b}$ de b fourni par la méthode du max. de vraisemblance n'est autre que

   $$\hat{b}(x_1, x_2, \ldots, x_n) = \max(x_1; x_2; \ldots; x_n)$$

   La fonction de répartition de la variable $M_n = \max(X_1; X_2; \ldots; X_n)$ est telle que

   $$\forall x \in [0; b], \quad F_{M_n}(x) = \mathbb{P}_b\{M_n \leq x\} = \mathbb{P}_b\{\max(X_1; X_2; \ldots; X_n) \leq x\} = \left(\frac{x}{b}\right)^n$$

   Le biais de l'estimateur $\hat{b}$ de b peut donc être calculé, on a

   $$\begin{aligned}
   \text{Biais}_{\hat{b}}(b) &= \mathbb{E}_b\left[\hat{b}(X_1, X_2, \ldots, X_n)\right] - b \\
   &= \int_0^b x \cdot f_{M_n}(x)dx - b \\
   &= \int_0^b x \cdot F'_{M_n}(x)dx - b \\
   &= \int_0^b x \cdot n\left(\tfrac{x}{b}\right)^{n-1}\tfrac{1}{b}dx - b \\
   &= \tfrac{n}{b^n}\int_0^b x^n dx - b \\
   &= \tfrac{n}{b^n}\left[\tfrac{x^{n+1}}{n+1}\right]_0^b - b \\
   &= \left(\tfrac{n}{b^n} \cdot \tfrac{b^{n+1}}{n+1}\right) - b
   \end{aligned}$$

On est donc en présence d'un "léger biais" $\frac{-b}{n+1}$, biais qui se fait de plus en plus discret à mesure que la taille n de l'échantillon considéré grandit ... Dans la pratique, on préférera cependant "prendre les devants" en corrigeant ce biais par une utilisation de l'estimateur

$$\tilde{b}(x_1, x_2, \ldots, x_n) = \frac{n+1}{n}\hat{b}(x_1, x_2, \ldots, x_n) = \frac{n+1}{n}\max(x_1; x_2; \ldots; x_n)$$

2. En revanche, les estimateurs de λ et de μ obtenus dans le contexte poissonien puis dans le contexte normal par la méthode du max. de vraisemblance sont, quant à eux, sans biais. En effet, si $X_1, X_2, \ldots, X_n$ est un échantillon i.i.d. de variables ayant pour espérance mathématique m, on a

$$\begin{array}{rcl} \mathbb{E}\left[\overline{X}_n\right] & = & \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] \\ & = & \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[X_k\right] \text{ (linéarité de l'espérance)} \\ & = & \frac{1}{n}\sum_{k=1}^{n} m \\ & = & \frac{1}{n}\cdot(nm) \\ & = & m \end{array}$$

Dans le contexte poissonien, on obtient donc

$$\text{Biais}_{\hat{\lambda}}(\lambda) = \mathbb{E}_\lambda\left[\hat{\lambda}(X_1, X_2, \ldots, X_n)\right] - \lambda = \mathbb{E}_\lambda\left[\overline{X}_n\right] - \lambda = \mathbb{E}_\lambda\left[X_1\right] - \lambda = 0,$$

et dans celui des variables $\mathcal{N}(\mu; 1)$, similairement :

$$\text{Biais}_{\hat{\mu}}(\mu) = \mathbb{E}_\mu\left[\hat{\mu}(X_1, X_2, \ldots, X_n)\right] - \mu = \mathbb{E}_\mu\left[\overline{X}_n\right] - \mu = \mathbb{E}_\mu\left[X_1\right] - \mu = 0$$

5.2 Intermezzo : Loi des Grands Nombres et TCL

En statistique inférentielle, comme on vient de le voir, l'estimation paramétrique requiert souvent de bien comprendre le comportement de la moyenne empirique

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$

associée à un échantillon i.i.d. de variables aléatoires.
Dans tout ce paragraphe, on supposera que ces variables i.i.d. $X_1, X_2, \ldots, X_n$ ont pour espérance (commune) un nombre réel m, et pour variance (commune) un réel strictement positif qui sera noté σ^2.
La ***Loi des Grands Nombres*** fournit une première information très naturelle au sujet du comportement asymptotique de $\overline{X}_n$: à mesure que $n \to +\infty$, les valeurs de la moyenne empirique $\overline{X}_n$ ont vocation à se concentrer à proximité de l'espérance mathématique m.
On dit aussi que les moyennes empiriques $\overline{X}_n$ convergent vers la "moyenne théorique" m.
Comment exprimer ce phénomène de convergence en respectant tout à fait l'exigence de rigueur mathématique ?

D'un point de vue mathématique, on peut affirmer que les moyennes $\overline{X}_n$ convergent ***en probabilité*** vers l'espérance mathématique m, ce qui signifie précisément que

$$\forall \epsilon > 0, \quad \mathbb{P}\left\{|\overline{X}_n - m| > \epsilon\right\} \underset{n\to\infty}{\longrightarrow} 0$$

En tout état de cause, lorsque la taille n de l'échantillon est suffisamment grande, on pourra considérer que $\overline{X}_n \approx m$, autrement dit que $\left(\overline{X}_n - m\right) \approx 0$. Mais quel est alors l'ordre de grandeur des fluctuations de $\overline{X}_n$ autour de m ? Observons que

$$\mathbb{E}\left[\overline{X}_n - m\right] = 0,$$

d'après la linéarité de l'espérance, tandis que

$$\begin{array}{rcl}
\mathrm{Var}\left(\overline{X}_n - m\right) & = & \mathrm{Var}\left(\overline{X}_n\right) \\
& = & \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \\
& = & \frac{1}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) \text{ (la variance est quadratique)} \\
& = & \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}\left(X_k\right) \text{ (les variables } X_1, X_2, \ldots, X_n \text{ sont indép.)} \\
& = & \frac{1}{n^2}\cdot n\sigma^2 \\
& = & \frac{\sigma^2}{n}
\end{array}$$

Ces fluctuations de $\overline{X}_n$ autour de l'espérance m sont donc de plus en plus petites, ayant pour ordre de grandeur $\frac{\sigma}{\sqrt{n}}$, l'écart-type de $\left(\overline{X}_n - m\right)$.
Le ***Théorème Central Limite*** (T.C.L.) affirme au sujet de ces fluctuations que la loi de $\left(\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}\right)$ devient très proche de celle d'une variable $\mathcal{N}(0;1)$ lorsque $n \to +\infty$.
Dans la pratique, cela signifie que pour évaluer la probabilité $\mathbb{P}\left\{a \leq \overline{X}_n \leq b\right\}$ de voir $\overline{X}_n$ atterrir dans l'intervalle $[a;b]$, on pourra ***centrer*** puis ***réduire*** $\overline{X}_n$, i.e. ***passer de*** $\overline{X}_n$ ***à*** $\left(\overline{X}_n - m\right)$ ***puis à*** $\left(\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}\right) = \tilde{X}_n$, avant de ***remplacer*** $\tilde{X}_n$ ***par une variable normale standard*** Z :

$$\begin{array}{rcl}
\mathbb{P}\left\{a \leq \overline{X}_n \leq b\right\} & = & \mathbb{P}\left\{(a-m) \leq (\overline{X}_n - m) \leq (b-m)\right\} \\
& = & \mathbb{P}\left\{\frac{\sqrt{n}}{\sigma}\cdot(a-m) \leq \frac{\sqrt{n}}{\sigma}\cdot(\overline{X}_n - m) \leq \frac{\sqrt{n}}{\sigma}\cdot(b-m)\right\} \\
& \approx & \mathbb{P}\left\{\frac{\sqrt{n}}{\sigma}\cdot(a-m) \leq Z \leq \frac{\sqrt{n}}{\sigma}\cdot(b-m)\right\} \\
& = & \mathbb{P}\left\{Z \leq \frac{\sqrt{n}}{\sigma}\cdot(b-m)\right\} - \mathbb{P}\left\{Z < \frac{\sqrt{n}}{\sigma}\cdot(a-m)\right\} \\
& = & F_Z(\frac{\sqrt{n}(b-m)}{\sigma}) - F_Z(\frac{\sqrt{n}(a-m)}{\sigma}),
\end{array}$$

tout du moins si n est suffisamment grand.
Le ***T.C.L.*** s'emploie tout aussi bien pour la somme $S_n = \sum_{k=1}^{n} X_k = n\overline{X}_n$: puisque son espérance vaut

$$\mathbb{E}\left[S_n\right] = \mathbb{E}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n}\mathbb{E}\left[X_k\right] = n \cdot m$$

et que sa variance vaut

$$\mathrm{Var}\,(S_n) = \mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathrm{Var}\,(X_k) = n \cdot \sigma^2,$$

centrer puis réduire S_n revient à remplacer cette variable-somme par $\tilde{S}_n = \frac{S_n - n\cdot m}{\sigma\sqrt{n}}$.
Par construction :

$$\mathbb{E}\left[\tilde{S}_n\right] = 0, \quad \mathrm{Var}\left(\tilde{S}_n\right) = 1,$$

et le T.C.L. affirme que la loi de $\tilde{S}_n$ est très proche de la loi $\mathcal{N}(0;1)$, tout du moins si n est suffisamment grand.
Dans ces conditions, on aura donc

$$\begin{array}{rcl}
\mathbb{P}\{a \leq S_n \leq b\} & = & \mathbb{P}\{(a-nm) \leq (S_n - nm) \leq (b-nm)\} \\
& = & \mathbb{P}\left\{\frac{1}{\sigma\sqrt{n}}\cdot(a-nm) \leq \frac{1}{\sigma\sqrt{n}}\cdot(S_n-nm) \leq \frac{1}{\sigma\sqrt{n}}\cdot(b-nm)\right\} \\
& \approx & \mathbb{P}\left\{\frac{1}{\sigma\sqrt{n}}\cdot(a-nm) \leq Z \leq \frac{1}{\sigma\sqrt{n}}\cdot(b-nm)\right\} \\
& = & \mathbb{P}\left\{Z \leq \frac{1}{\sigma\sqrt{n}}\cdot(b-nm)\right\} - \mathbb{P}\left\{Z < \frac{1}{\sigma\sqrt{n}}\cdot(a-nm)\right\} \\
& = & F_Z(\frac{1}{\sigma\sqrt{n}}\cdot(b-nm)) - F_Z(\frac{1}{\sigma\sqrt{n}}\cdot(a-nm))
\end{array}$$

A ce stade, l'évaluation de la fonction F_Z se fait en recourant à une table de valeurs numériques pour la répartition normale standard.
Voyons tout de suite un exemple fondamental d'application du T.C.L., exemple qui manque rarement de surprendre le novice et constitue une bonne histoire à retenir !
Exemple d'Application du T.C.L. dans un contexte binomial :
Les candidats BO et NG s'affrontent pour obtenir les faveurs d'un corps électoral comportant un million de votants.
Le jour de l'élection, sur tous ces votants, 1000 seulement ont pris une décision claire, celle de voter pour BO. Quant aux 999'000 autres, ils sont totalement indécis et choisissent de remplir leur devoir civique en tirant à Pile ou Face dans l'isoloir pour voter ensuite en faveur de BO ou de NG.
Quelle est la loi de la variable X comptant le nombre de bulletins favorables à NG ?
Comment estimer la probabilité de voir BO remporter ces élections ?

Partons du principe que la pièce utilisée par les électeurs indécis est équilibrée, et désignons par $X_1, X_2, \ldots, X_n$ les variables représentant le vote de chacun de ces $n = 999'000$ électeurs. On a

$$X = (X_1 + X_2 + \ldots + X_n) = \sum_{k=1}^{n} X_k,$$

la variable X comptant le nombre de bulletins favorables à NG suit donc une loi binomiale de paramètres $n = 999'000$ et $p = \frac{1}{2}$.

Mathématiquement, l'événement *"E: BO remporte les élections"* peut être décrit comme suit :

$$E = \{X < 500'000\} = \{\omega \in \Omega | \, X(\omega) < 500'000\}$$

On a donc

$$\begin{array}{rcl} \mathbb{P}\{E\} & = & \mathbb{P}\{X < 500'000\} \\ & = & \sum_{k=0}^{499'000} \mathbb{P}\{X = k\} \\ & = & \sum_{k=0}^{499'000} \binom{999'000}{k} \frac{1}{2}^k (1-\frac{1}{2})^{999'000-k} \\ & = & \left(\frac{1}{2}\right)^{999'000} \sum_{k=0}^{499'000} \binom{999'000}{k} \end{array}$$

Cette dernière expression paraît tout simplement impossible à évaluer.
N'oublions pas cependant que le *Théorème Central Limite* s'applique à X, qui est une *vaste somme de* $999'000$ *variables i.i.d.* ! On obtient ainsi :

$$\begin{array}{rcl} \mathbb{P}\{E\} & = & \mathbb{P}\{X < 500'000\} \\ & = & \mathbb{P}\left\{\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} < \frac{500'000-\mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}\right\} \\ & = & \mathbb{P}\left\{\frac{X-499'500}{\sqrt{249'750}} < \frac{500'000-499'500}{\sqrt{249'750}}\right\} \\ & = & \mathbb{P}\left\{\frac{X-499'500}{\sqrt{249'750}} < 1\right\} \\ & \approx & \mathbb{P}\{Z < 1\}, \end{array}$$

où Z désigne à nouveau une variable normale standard.
En faisant usage de la table des valeurs numériques de la répartition normale standard, on trouve ensuite que

$$\mathbb{P}\{Z < 1\} = F_Z(1) \cong 0,8413.$$

Il serait donc tout à fait étonnant que BO ne remporte pas ces élections, puisque la probabilité

$$\mathbb{P}\{\text{'BO remporte ces élections'}\} = \mathbb{P}\{X < 500'000\}$$

s'élève à plus de 84% !

5.3 Intervalles de Confiance

Prenons à present un autre point de vue pour estimer le paramètre d'une loi de variable aléatoire : au lieu de proposer une valeur précise pour le paramètre, on va situer ce paramètre dans un *intervalle* correspondant à un certain *niveau de confiance*.

1er Cas de Figure : Moyenne d'un échantillon gaussien à variance connue

Supposons que $X_1, X_2, \ldots, X_n$ est un échantillon i.i.d. de variables $\mathcal{N}(\mu; 1)$ de variables normales (on dit aussi *gaussiennes*) d'espérance μ inconnue et de variance $\sigma^2 = 1$. On a enregistré les résultats suivants :

$$X_1 = x_1, \, X_2 = x_2, \, \ldots, \, X_n = x_n,$$

en sorte que $\mu \approx \overline{x}_n$, par la méthode du max.de vraisemblance.
Comment régler la valeur de $\varepsilon > 0$ pour être sûr à 95% que la véritable valeur de μ est située

dans l'intervalle $[\overline{x}_n - \varepsilon; \overline{x}_n + \varepsilon]$?
Observons que

$$\mathbb{P}_\mu\left\{\mu - \varepsilon \leq \overline{X}_n \leq \mu + \varepsilon\right\} = \mathbb{P}_\mu\left\{\overline{X}_n \in [\mu - \varepsilon; \mu + \varepsilon]\right\} = \mathbb{P}_\mu\left\{[\overline{X}_n - \varepsilon; \overline{X}_n + \varepsilon] \ni \mu\right\},$$

il s'agit donc de régler $\varepsilon > 0$ en sorte que $\mathbb{P}_\mu\left\{\mu - \varepsilon \leq \overline{X}_n \leq \mu + \varepsilon\right\} = 0,95$.
Pour avancer dans l'élaboration de cet intervalle de confiance, nous aurons besoin d'utiliser une propriété essentielle de ***stabilité des variables gaussiennes*** : les variables

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,\ \left(\overline{X}_n - \mu\right),\ \sqrt{n}\left(\overline{X}_n - \mu\right),$$

construites linéairement à partir de la somme $(\sum_{k=1}^{n} X_k)$, sont encore des variables gaussiennes, avec

$$\mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] = \mu, \quad \mathbb{E}\left[\overline{X}_n - \mu\right] = 0 = \mathbb{E}\left[\sqrt{n}\left(\overline{X}_n - \mu\right)\right]$$

et

$$\mathrm{Var}\left[\sqrt{n}\left(\overline{X}_n - \mu\right)\right] = n \cdot \mathrm{Var}\left[\overline{X}_n - \mu\right] = n \cdot \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}\left[X_k\right] = 1$$

$Z = \sqrt{n}\left(\overline{X}_n - \mu\right)$ est donc une variable normale standard, et l'on a ensuite

$$\begin{array}{rcl}\mathbb{P}_\mu\left\{\mu - \varepsilon \leq \overline{X}_n \leq \mu + \varepsilon\right\} & = & \mathbb{P}_\mu\left\{-\varepsilon \leq \overline{X}_n - \mu \leq \varepsilon\right\} \\ & = & \mathbb{P}_\mu\left\{-\sqrt{n}\varepsilon \leq \sqrt{n}(\overline{X}_n - \mu) \leq \sqrt{n}\varepsilon\right\} \\ & = & \mathbb{P}_\mu\left\{-\sqrt{n}\varepsilon \leq Z \leq \sqrt{n}\varepsilon\right\} \\ & = & F_Z(\sqrt{n}\varepsilon) - F_Z(-\sqrt{n}\varepsilon)\end{array}$$

L'utilisation d'une table de quantiles asociés à la loi normale standard permet alors de constater que l'on atteint un niveau de confiance situé à 95% lorsque $\sqrt{n}\varepsilon = \varepsilon_{0,05} = 1,96$, ce qui permet de régler la valeur de ε à $\frac{1,96}{\sqrt{n}}$.
L'intervalle de confiance au niveau 95% pour le paramètre de moyenne μ est donc

$$I = [\overline{x}_n - \frac{1,96}{\sqrt{n}}; \overline{x}_n + \frac{1,96}{\sqrt{n}}]$$

Remarquons que

- Si l'on souhaite obtenir un intervalle de confiance à un autre niveau $(1-\alpha)$, il suffit de régler la valeur de ε à $\frac{\varepsilon_\alpha}{\sqrt{n}}$, l'intervalle de confiance au niveau $(1-\alpha)$ pour le paramètre de moyenne μ est donc

$$I = [\overline{x}_n - \frac{\varepsilon_\alpha}{\sqrt{n}}; \overline{x}_n + \frac{\varepsilon_\alpha}{\sqrt{n}}],$$

la valeur numérique de ε_α tel que $F_Z(-\varepsilon_\alpha) = 1 - F_Z(\varepsilon_\alpha) = \frac{\alpha}{2}$ pouvant être obtenue directement dans notre Table de référence.
Par exemple, l'intervalle de confiance au niveau 99% pour le paramètre μ est

$$I' = [\overline{x}_n - \frac{2,576}{\sqrt{n}}; \overline{x}_n + \frac{2,576}{\sqrt{n}}]$$

(on a "élargi" l'intervalle précédent afin d'augmenter le niveau de confiance).

- Si la variance commune à toutes les variables de notre échantillon gaussien n'est plus 1 mais une autre grandeur connue $\sigma^2 > 0$, l'intervalle de confiance au niveau $(1-\alpha)$ pour le paramètre μ est donné par

$$I = [\overline{x}_n - \frac{\sigma \times \varepsilon_\alpha}{\sqrt{n}}; \overline{x}_n + \frac{\sigma \times \varepsilon_\alpha}{\sqrt{n}}]$$

(plus la variance commune aux variables i.i.d. de l'échantillon est grande, plus l'intervalle de confiance est large).

Exemple 1 :
on suppose que le taux de présence d'un gaz nocif dans l'atmosphère, mesuré en $\mu g/m^3$, suit une loi normale d'espérance τ inconnue et d'écart-type $\sigma = 10\mu g/m^3$. On effectue tout d'abord dix mesures de ce taux, ce qui permet d'obtenir des valeurs $x_1, x_2, \ldots, x_{10}$ telles que $\sum_{i=1}^{10} x_i = 504 g/m^3$. Une seconde équipe effectue une centaine de mesures et obtient des valeurs $x'_1, x'_2, \ldots, x'_{100}$ telles que $\sum_{i=1}^{100} x'_i = 5065 g/m^3$; enfin, une troisième effectue 400 mesures, obtenant des valeurs $x''_1, x''_2, \ldots, x''_{400}$ telles que $\sum_{i=1}^{400} x''_i = 20208 g/m^3$.
Dans ces conditions, quels sont les intervalles de confiance au niveau 90% obtenus par chacune des trois équipes pour le paramètre τ ?
On a, pour la première équipe : $\overline{x}_n = \overline{x}_{10} = \frac{1}{10}\sum_{i=1}^{10} x_i = 50,4 g/m^3$, et l'intervalle de confiance recherché au niveau $(1-\alpha)$ est donc de la forme

$$I = [\overline{x}_n - \varepsilon; \overline{x}_n + \varepsilon] = [50,4 - \varepsilon; 50,4 + \varepsilon],$$

la constante positive ε devant être réglée en sorte que

$$\mathbb{P}\left\{\tau - \varepsilon \leq \overline{X}_n \leq \tau + \varepsilon\right\} = (1-\alpha)$$

En utilisant le fait que $\overline{X}_n$ suit une loi $\mathcal{N}(\tau; \frac{\sigma^2}{n} = 10)$, on obtient par exemple au niveau 90% :

$$\mathbb{P}\left\{\tau - \varepsilon \leq \overline{X}_n \leq \tau + \varepsilon\right\} = \mathbb{P}\left\{\frac{-\varepsilon}{\sqrt{10}} \leq \frac{(\overline{X}_n - \tau)}{\sqrt{10}} \leq \frac{\varepsilon}{\sqrt{10}}\right\} = \mathbb{P}\left\{\frac{-\varepsilon}{\sqrt{10}} \leq Z \leq \frac{\varepsilon}{\sqrt{10}}\right\} = 2F_Z(\frac{\varepsilon}{\sqrt{10}}) - 1,$$

il nous faut donc $\frac{\varepsilon}{\sqrt{10}} = 1,65$ au niveau $(1-\alpha) = 90\%$, ce qui nous donne l'intervalle de confiance

$$I = [50,4 - \sqrt{10} \cdot 1,65; 50,4 + \sqrt{10} \cdot 1,65] = [50,4 - 5,22; 50,4 + 5,22] = [45,18; 55,62]$$

Si la deuxième équipe cherche à fabriquer un intervalle de confiance I' pour ce même niveau de 90%, il lui faut utiliser le fait que la moyenne empirique $\overline{X}_{100}$ suit une loi $\mathcal{N}(\tau; \frac{\sigma^2}{n'} = \frac{10^2}{100} = 1)$, ce qui permet d'écrire

$$\mathbb{P}\left\{\tau - \varepsilon' \leq \overline{X}_{n'} \leq \tau + \varepsilon'\right\} = \mathbb{P}\left\{-\varepsilon' \leq (\overline{X}_{n'} - \tau) \leq \varepsilon'\right\} = F_Z(\varepsilon') - F_Z(-\varepsilon') = 2F_Z(\varepsilon') - 1,$$

pour $n' = 100$ ce nouvel intervalle de confiance au niveau 90% est donc

$$I' = [50,4 - 1,65; 50,4 + 1,65] = [48,75; 52,05]$$

Enfin pour la troisième équipe, toujours au niveau 90%, en utilisant le fait que $\overline{X}_{n''}$ suit une loi $\mathcal{N}(\tau; \frac{\sigma^2}{n''} = \frac{10^2}{400} = 0,25)$ on obtient

$$\mathbb{P}\left\{\tau - \varepsilon'' \leq \overline{X}_n \leq \tau + \varepsilon''\right\} = \mathbb{P}\left\{\frac{-\varepsilon''}{\sqrt{0,25}} \leq \frac{(\overline{X}_n - \tau)}{\sqrt{0,25}} \leq \frac{\varepsilon''}{\sqrt{0,25}}\right\} = 2F_Z(\frac{\varepsilon''}{\sqrt{0,25}}) - 1,$$

il nous faut donc $\frac{\varepsilon''}{\sqrt{0,25}} = 2\varepsilon = 1,65$ au niveau $(1-\alpha) = 90\%$, ce qui nous donne l'intervalle de confiance

$$I'' = [50,4 - \frac{1,65}{2}; 50,4 + \frac{1,65}{2}] = [49,575; 51,225]$$

Alternativement, au niveau de confiance 95%, ces intervalles sont rendus plus larges :

$I = [50,4 - \sqrt{10}\cdot 1,96; 50,4 + \sqrt{10}\cdot 1,96] = [44,202; 56,598]$, $I' = [50,4 - 1,96; 50,4 + 1,96] = [48,44; 52,36]$, $I'' = [50,4 - \frac{1,96}{2}; 50,4 + \frac{1,96}{2}] = [49,42; 51,38]$

(car $F_Z(1,96) \simeq 0,975$ si Z suit une loi $\mathcal{N}(0;1)$).
Enfin, au niveau de confiance 99% :

$I = [50,4 - \sqrt{10}\cdot 2,576; 50,4 + \sqrt{10}\cdot 2,576] = [42,254; 58,546]$, $I' = [47,824; 52,976]$, $I'' = [50,4 - \frac{2,576}{2}; 50,4 + \frac{2,576}{2}] = [49,112; 51,688]$

(car $F_Z(2,576) \simeq 0,995$ si Z suit une loi $\mathcal{N}(0;1)$).

2ème Cas de Figure : Moyenne d'un échantillon gaussien à variance inconnue

Supposons que $X_1, X_2, \ldots, X_n$ est un échantillon i.i.d. de variables $\mathcal{N}(\mu; \sigma^2)$ de variables normales d'espérance μ et de variance $\sigma^2 = 1$ *toutes deux inconnues*. On a enregistré les résultats suivants :

$$X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n,$$

en sorte que $\mu \approx \overline{x}_n$, par la méthode du max.de vraisemblance.
Dans ces conditions, comment régler la valeur de $\varepsilon > 0$ pour être sûr à 95% que la véritable valeur de μ est située dans l'intervalle $[\overline{x}_n - \varepsilon; \overline{x}_n + \varepsilon]$?
L'idée est alors de commencer par estimer la valeur de la variance σ^2 commune aux variables considérées.
Il s'avère que l'estimateur le plus naturel de σ^2, à savoir l'estimateur du max. de vraisemblance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2$$

comporte un biais :

$$\mathbb{E}\left[\hat{\sigma}^2(X_1, X_2, \ldots, X_n)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] = \frac{n-1}{n}\sigma^2$$

On corrige donc ce biais en utilisant plutôt l'estimateur s^2 de σ^2 défini par

$$s^2 = s(x_1, x_2, \ldots, x_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\cdot\overline{x}_n^2$$

Avec cette estimation de σ^2 en poche, on raisonne ensuite exactement comme dans le contexte du paragraphe précédent, à ceci près que la variable

$$\frac{\overline{X}_n - \mu}{s(X_1, X_2, \ldots, X_n)/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\cdot\overline{X}_n^2}}$$

ne suit plus *exactement* une loi gaussienne : pour n "petit", cette variable suit une *loi de Student* à $\nu = (n-1)$ degrés de liberté.
Cependant, lorsque n est grand, une telle loi de Student est *extrêmement proche de la loi normale standard* ! On retrouve alors la table de quantiles utilisée précédemment. En effet :

$$\begin{aligned}
\mathbb{P}_\mu\left\{\mu - \varepsilon \leq \overline{X}_n \leq \mu + \varepsilon\right\} &= \mathbb{P}_\mu\left\{-\varepsilon \leq \overline{X}_n - \mu \leq \varepsilon\right\} \\
&= \mathbb{P}_\mu\left\{-\tfrac{\sqrt{n}}{s}\varepsilon \leq \sqrt{n}(\overline{X}_n - \mu) \leq \tfrac{\sqrt{n}}{s}\varepsilon\right\} \\
&= \mathbb{P}_\mu\left\{-\tfrac{\sqrt{n}}{s}\varepsilon \leq Z \leq \tfrac{\sqrt{n}}{s}\varepsilon\right\} \\
&= F_Z(\tfrac{\sqrt{n}}{s}\varepsilon) - F_Z(-\tfrac{\sqrt{n}}{s}\varepsilon)
\end{aligned}$$

Si la taille n de l'échantillon considéré est suffisamment grande ($n \geq 50$), il convient donc de régler la valeur de ε à $\frac{s}{\sqrt{n}}\cdot\varepsilon_\alpha$, où ε_α est encore tel que $F_Z(\varepsilon_\alpha) = 1 - \frac{\alpha}{2}$.
Par exemple, l'intervalle de confiance I obtenu au niveau 95% est donné par

$$I = [\overline{x}_n - \frac{s}{\sqrt{n}}\cdot\varepsilon_\alpha; \overline{x}_n + \frac{s}{\sqrt{n}}\cdot\varepsilon_\alpha] = \left[\overline{x}_n - \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}}{\sqrt{n}}\cdot 1,96; \overline{x}_n + \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}}{\sqrt{n}}\cdot 1,96\right],$$

tandis qu'au niveau 99% on obtient l'intervalle

$$J = \left[\overline{x}_n - \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}}{\sqrt{n}}\cdot 2,576; \overline{x}_n + \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}}{\sqrt{n}}\cdot 2,576\right]$$

Exemple 2 :
on suppose maintenant que le taux de présence d'un gaz nocif dans l'atmosphère, mesuré en

$\mu g/m^3$, suit une loi normale d'espérance τ inconnue et d'écart-type σ *lui aussi inconnu*. On effectue tout d'abord dix mesures de ce taux, ce qui permet d'obtenir des valeurs $x_1, x_2, \ldots, x_{10}$ telles que $\sum_{i=1}^{10} x_i = 504 g/m^3$ et $\sum_{i=1}^{10} x_i = 26'425\,(g/m^3)^2$.
Une seconde équipe effectue une centaine de mesures et obtient des valeurs $x'_1, x'_2, \ldots, x'_{100}$ telles que $\sum_{i=1}^{100} x'_i = 5065 g/m^3$ et $\sum_{i=1}^{100} (x'_i)^2 = 268'742\,(g/m^3)^2$; enfin, une troisième effectue 400 mesures, obtenant des valeurs $x''_1, x''_2, \ldots, x''_{400}$ telles que $\sum_{i=1}^{400} x''_i = 20208 g/m^3$ et $\sum_{i=1}^{100} (x''_i)^2 = 1'068'200\,(g/m^3)^2$.
Dans ces conditions, quels sont les intervalles de confiance au niveau 90% obtenus par chacune des trois équipes pour le paramètre τ ?
On a, pour la première équipe : $\overline{x}_n = \overline{x}_{10} = \frac{1}{10}\sum_{i=1}^{10} x_i = 50,4 g/m^3$, mais aussi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\cdot \overline{x}_n^2 = \frac{1}{9}\cdot 26'425 - \frac{10}{9}\cdot (50,4)^2 = 113,71$$

L'intervalle de confiance recherché au niveau $(1-\alpha) = 90\%$ est donc de la forme

$$I = [\overline{x}_n - \frac{s}{\sqrt{n}}\cdot t_\alpha ; \overline{x}_n + \frac{s}{\sqrt{n}}\cdot t_\alpha]$$

pour $\overline{x}_n = 50,4$, $s = \sqrt{113,71} \simeq 10,66$, $\sqrt{n} = \sqrt{10} \simeq 3,16$ et $t_\alpha = 1,833$, t_α désignant ici le quantile d'une variable de Student T à 9 deg. de liberté, en sorte que : $\mathbb{P}(|T| > t_\alpha) = 1-\alpha = 0,1$.
Ainsi :

$$I = [50,4 - \frac{10,66}{3,16}\cdot 1,833 ; 50,4 + \frac{10,66}{3,16}\cdot 1,833] = [44,22 ; 56,58]$$

Pour la deuxième équipe, puisque

$$\overline{x'} = \frac{1}{100}\sum_{i=1}^{100} x'_i = 50,65 g/m^3,\ (s')^2 = \frac{1}{99}\cdot 268'742 - \frac{100}{99}\cdot (50,65)^2 \simeq 123,23,$$

on obtient, toujours au niveau de confiance $(1-\alpha) = 90\%$:

$$I' = [\overline{x'} - \frac{s'}{\sqrt{n'}}\cdot \varepsilon_\alpha ; \overline{x'} + \frac{s'}{\sqrt{n'}}\cdot \varepsilon_\alpha] = [50,65 - \frac{11,1}{10}\cdot 1,65 ; 50,65 + \frac{11,1}{10}\cdot 1,65] = [48,82 ; 52,48]$$

où $\varepsilon_\alpha = 1,65$ désigne un quantile de variable $\mathcal{N}(0;1)$. Enfin, pour la troisième équipe, puisque

$$\overline{x''} = \frac{1}{400}\sum_{i=1}^{400} x''_i = 50,52 g/m^3,\ (s'')^2 = \frac{1}{399}\cdot 1'068'200 - \frac{400}{399}\cdot (50,52)^2 \simeq 118,53,$$

l'intervalle de confiance au niveau 90% est l'intervalle

$$I'' = [\overline{x''} - \frac{s''}{\sqrt{n''}}\cdot \varepsilon_\alpha ; \overline{x''} + \frac{s''}{\sqrt{n''}}\cdot \varepsilon_\alpha] = [50,52 - \frac{10,89}{20}\cdot 1,65 ; 50,52 + \frac{10,89}{20}\cdot 1,65] = [49,62 ; 51,42]$$

3ème Cas de Figure : Paramètre p commun à des variables de Bernoulli
Plaçons-nous maintenant dans une situation très couramment rencontrée dans la pratique : on suppose que $X_1, X_2, \ldots, X_n$ sont des variables de Bernoulli indépendantes de paramètre p inconnu, et l'on cherche à estimer p. Ayant enregistré les résultats

$$X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n \ \textit{(où } x_1, x_2, \ldots, x_n \in \{0;1\}\textit{)}$$

on dispose d'un estimateur naturel de p par la méthode du max.de vraisemblance : cet estimateur $\hat{p}$ est tout simplement donné par $\hat{p} = \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.
A partir de cette estimation ponctuelle, comment régler la valeur du nombre positif ε pour faire en sorte que l'intervalle

$$I = [\hat{p} - \varepsilon; \hat{p} + \varepsilon] = [\overline{x}_n - \varepsilon; \overline{x}_n + \varepsilon]$$

recouvre la vraie valeur de p à un certain niveau de confiance $(1-\alpha)$?
Tout comme dans le paragraphe précédent, il s'agira de commencer par estimer l'écart-type $\sigma = p(1-p)$ de chacune des variables de Bernoulli $X_1, X_2, \ldots, X_n$. A l'évidence, le plus naturel sera de considérer que

$$\sigma \approx \hat{p}(1-\hat{p})$$

D'après le TCL, si n est grand $(n \geq 50)$ on pourra considérer que la variable

$$\frac{\overline{X}_n - p}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \approx \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\hat{p}(1-\hat{p})}}$$

suit, à très peu de choses près, une loi normale standard. On aura donc

$$\begin{aligned}
\mathbb{P}\left\{\overline{X}_n - \varepsilon \leq p \leq \overline{X}_n + \varepsilon\right\} &= \mathbb{P}\left\{-\varepsilon \leq \overline{X}_n - p \leq \varepsilon\right\} \\
&= \mathbb{P}\left\{-\frac{\varepsilon\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \leq \frac{(\overline{X}_n - p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \leq \frac{\varepsilon\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}\right\} \\
&= 2F_Z\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}\right) - 1\,,
\end{aligned}$$

il convient donc de régler la constante $\varepsilon > 0$ en sorte que

$$F_Z\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}}\right) = 1 - \frac{\alpha}{2}, \text{ soit } \varepsilon = \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\,.$$

L'intervalle de confiance obtenu au niveau $(1-\alpha)$ pour la vraie valeur de p est donc

$$I = [\hat{p} - \varepsilon; \hat{p} + \varepsilon] = [\hat{p} - \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}; \hat{p} + \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}]$$

Exemple 3 :
Dans un grand pays démocratique, un quotidien publie chaque mois la "cote" du chef du gouvernement à partir des résultats d'un sondage effectué sur un échantillon représentatif de $n = 1000$ personnes.

Les sondages de janvier et février ont donné respectivement les cotes de 38% et 36%.
Dans ces conditions, quels sont les intervalles de confiance I_J et I_F contenant les véritables cotes p_J puis p_F correspondant à ces deux mois, au niveau de confiance 95% ? Peut-on raisonnablement affirmer que cette cote de popularité a chuté de deux points entre janvier et février ??

Puisque ce niveau de confiance correspond au quantile gaussien standard $\varepsilon_\alpha = 1,96$, l'application des formules vues précédemment donne pour le mois de janvier

$$\begin{aligned} I_J &= [\hat{p}_J - \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}_J(1-\hat{p}_J)}}{\sqrt{n}}; \hat{p}_J + \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}_J(1-\hat{p}_J)}}{\sqrt{n}}] \\ &= [0,38 - 1,96 \cdot \frac{\sqrt{0,38(1-0,38)}}{\sqrt{1000}}; 0,38 + 1,96 \cdot \frac{\sqrt{0,38(1-0,38)}}{\sqrt{1000}}] \\ &= [0,35; 0,41] \end{aligned}$$

puis pour le mois de février :

$$\begin{aligned} I_F &= [\hat{p}_F - \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}_F(1-\hat{p}_F)}}{\sqrt{n}}; \hat{p}_F + \varepsilon_\alpha \cdot \frac{\sqrt{\hat{p}_F(1-\hat{p}_F)}}{\sqrt{n}}] \\ &= [0,36 - 1,96 \cdot \frac{\sqrt{0,36(1-0,36)}}{\sqrt{1000}}; 0,36 + 1,96 \cdot \frac{\sqrt{0,36(1-0,36)}}{\sqrt{1000}}] \\ &= [0,33; 0,39] \end{aligned}$$

Ces deux intervalles de confiance se chevauchent largement, il serait donc donc exagéré d'écrire dans la chronique politique du mois de février que la cote du chef de gouvernement vient de perdre 2 points !

5.4 Exercices d'Application

I Intervalles de confiance et fiabilité des sondages :

On considère un échantillon de 1000 personnes auprès desquelles on effectue un sondage avant un référendum : 450 répondent *oui*.

1) Quelle valeur proposer pour estimer la probabilité p, probabilité qu'un électeur vote *oui* le jour du scrutin ?

2) Calculer un intervalle de confiance, centré sur cette estimation, pour le paramètre p avec une confiance de 0,95 puis 0,98.

3) Proposer un intervalle de la forme $[0; \dots]$ dans lequel se trouve p avec une confiance de 0,95.

4) Quelle confiance accordez-vous à l'affirmation: "le *non* va l'emporter"?

5) Combien faut-il interroger de personnes pour obtenir une fourchette pour p de largeur $2 \times 0,01$?

II Arrondis et TCL :

Dans un supermarché, aprés avoir pesé une marchandise, la balance électronique délivre un ticket où le prix est arrondi au centime près.

1) Pour une pesée, on note X la variable aléatoire donnant, en centimes, la différence entre le prix indiqué sur l'étiquette et le prix exact. Expliquer pourquoi on peut modéliser la loi de X par une loi uniforme sur l'intervalle $[-0,5; +0,5]$. Calculer l'espérance et la variance de X.

2) Quelle est la perte moyenne du supermarché pour 10^4 unités pesées? Quelle est la probabilité que le supermarché gagne plus de 20 centimes pour 10^4 unités? perde plus de 50 euros pour 10^6 unités?

III Stratégies d'examen :

Un QCM comporte 100 questions. Deux réponses sont proposées pour chaque question, l'une d'entre elles est juste, l'autre est fausse. Un étudiant décide de répondre à chaque question en choisissant au hasard sa réponse.

1) Calculer la probabilité qu'il ait au moins 50 réponses justes.

2) Compléter les phrases " Avec une probabilité de 95%, le nombre de réponses justes est supérieur à ..." et "Avec 5% de chance, le nombre de réponses justes est supérieur à ..."

3) Le professeur hésite entre deux modes de notation. Le premier consiste à noter +1 les réponses justes, 0 sinon ; le second consiste à noter +1 les réponses justes, 0 les non-réponses et -1 les réponses fausses. On appelle Y et Z les notes (aléatoires) qu'obtiendra l'étudiant avec le premier mode et avec le second mode de calcul. Quelles notes moyennes l'étudiant peut-il espérer? A-t-il intérêt à changer de stratégie en choisissant entre "non-réponse", "réponse 1" ou "réponse 2" avec les probabilités respectives q, p et p ?

IV TCL et Mesures Physiques :

En poste à l'Observatoire de Göttingen, l'astronome *Carl-Friedrich G.* décide de mesurer la distance séparant son lieu de travail d'une étoile lointaine.
Connaissant fort bien le matériel dont il dispose et prenant en compte les aléas liés à chaque mesure (perturbations atmosphériques, etc...), il part du principe que chaque mesure M ne lui fournit qu'une valeur aléatoire d'espérance d (vraie valeur de cette distance, en années-lumière) et d'écart-type 2 années lumière.
Dans ces conditions, il décide de prendre un certain nombre de mesures *indépendantes* $M_1, M_2, \ldots, M_n$, puis d'adopter leur moyenne

$$\overline{M}_n = \frac{M_1 + M_2 + \ldots + M_n}{n}$$

comme évaluation de cette distance.
Combien de mesures indépendantes notre astronome doit-il effectuer pour être raisonnablement sûr (sûr à 95%) de ne pas se tromper au-delà d'une demi-année lumière dans son évaluation?

V TCL et garanties :

Un fabriquant de clefs USB sait qu'une proportion de 0,05 de sa production est défectueuse.
1) Il garantit à un client qui achète 10000 pièces de la rembourser si plus de m clefs sont défectueuses. Comment le fabriquant doit-il choisir m pour n'avoir pas plus d'une "chance" sur cent d'avoir à rembourser son client?
2) Le service des fraudes vient faire un contrôle, prélève 1000 clefs et note S le nombre de clefs défectueuses.
a) Donner un intervalle de la forme $[50 - a, 50 + a]$ dans lequel S prend ses valeurs avec une probabilité supérieure à 0,95.
b) Idem avec un intervalle de la forme $[0, x]$, puis $[y, \infty[$. Traduire les résultats par des phrases.

VI Surcharge :

Un avion long courrier peut transporter 100 passagers et leurs bagages. L'avion pèse 120 tonnes sans passagers ni bagages, mais équipage compris et le plein de carburant effectué. Les consignes de sécurité interdisent au commandant de bord de décoller si le poids de l'appareil chargé dépasse 129 tonnes. Les 100 places ont été réservées. Le poids d'un voyageur est une v.a. X d'espérance mathématique 70 kg et d' écart-type 10 kg. Le poids des bagages d'un voyageur est une v.a. Y d'espérance mathématique 20 kg et d'écart-type 10 kg. Toutes ces variables aléatoires sont mutuellement indépendantes, leurs lois de probabilité sont inconnues.
1) Soit C la v.a. égale au poids de l'avion lorsque les 100 voyageurs et leurs bagages seront chargés.

a- Donner l'expression en kilos de C en fonction des différentes v.a. décrites dans l'énoncé.
b- Calculer l'espérance et la variance de C et donner la loi de C.
c- Quelle est la probabilité p que le commandant refuse d'embarquer une partie des bagages afin de respecter les consignes de sécurité ?
2) L'avion doit effectuer 800 vols dans des conditions identiques à celles décrites ci-dessus. On fait toutes les hypothèses d'indépendance nécessaires. Soit N le nombre de vols où le commandant refuse d'embarquer une partie des bagages afin de respecter les consignes de sécurité.
a- Quelle est la loi de probabilité de N ? Quelle approximation peut-on en faire?
b- Quelle est la probabilité qu'au moins une fois parmi les 800 vols, le commandant refuse d'embarquer une partie des bagages ?

VII Sondages et Audimat :

On interroge 400 téléspectateurs choisis au hasard. Parmi eux, 152 individus déclarent avoir regardé la nouvelle émission *"Perdu en Translation"* diffusée le mardi soir, les 248 autres déclarent ne pas avoir regardé cette nouvelle et passionnante émission.

1. En donnant successivement à $(1-\alpha)$ les valeurs 90%, 95%, 98% puis 99%, estimer par un intervalle de confiance au niveau $(1-\alpha)$ le pourcentage de téléspectateurs ayant regardé cette émission.
2. Quelle aurait dû être la taille n de l'échantillon de téléspectateurs interrogés si le but avait été de déterminer ce pourcentage à 2% près, pour chacun des niveaux de confiance envisagés dans la question précédente ?
3. Tracer deux courbes correspondant aux bornes des intervalles de confiance au niveau 95% en fonction de n (on pourra donner à n les valeurs 100, 150, 200, 250, ..., 1000).

VIII Etre ou ne pas être ... élu :

A la veille d'une importante consultation électorale, on effectue un sondage sur un échantillon de $n = 100$ votants. Cet échantillon fournit 55 intentions de vote favorables au candidat A.

1. Donner un intervalle de confiance au niveau $0,99$ pour le pourcentage d'électeurs favorables au candidat A.
2. Quelle doit être la taille n de l'échantillon pour que la largeur de l'intervalle de confiance au niveau 99% soit au plus de 12% ?
3. Quelle devrait être la taille de l'échantillon fournissant 55% d'intentions de vote favorables au candidat A pour que ce candidat soit sûr au niveau de risque 1% d'être élu ??

IX Le TCL est universel ⊞ :

Choisir une loi parmi celles mentionnées dans l'exercice 13 et enregistrer dans la variable m la valeur de l'espérance et dans la variable s l'écart-type de cette loi.

1. Pour $n = 2$ puis $n = 10$ puis $n = 30$ puis $n = 100$, simuler n réalisations de cette loi (désignées par $X_1, \ldots, X_n$) et calculer la moyenne $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ de ces réalisations.

2. Reprendre 1000 fois la question 1) en stockant les 1000 valeurs des moyennes $\bar{X}_n$.

3. Centrer et réduire ces 1000 valeurs en effectuant le calcul de $(\bar{X}_n - m)/(s/\sqrt{n})$.

4. Regrouper les 1000 valeurs de la question précédentes en 10 classes centrées en 0 et de largeur 1, et tracer l'histogramme correspondant.

5. Comparer avec les histogrammes des autres étudiants du groupe. Que constatez-vous? (*effectuer les comparaisons pour les 4 valeurs proposées de* n). Décrire précisément la conséquence du TCL qui vient ainsi d'être illustrée.

X Planche de Galton ⊞ :

- Qu'est-ce qu'une planche de Galton (*Galton Board*, en anglais) ? [3]
- En quoi est-ce qu'une telle planche fournit une illustration du TCL ?
- Comment vous y prendriez-vous pour *simuler* une telle expérience ?

XI Calcul d'une valeur approchée de π ⊞ :

On se propose de calculer une valeur approchée de π par une méthode probabiliste, appelée méthode de Monte-Carlo. Il s'agit en fait d'une illustration de la loi des grands nombres et du théorème central limite.

Méthode : On lance un grand nombre de points dans le carré $[-1, 1] \times [-1, 1]$ et on compte la proportion de points qui se trouvent à l'intérieur du disque de centre $(0, 0)$ et de rayon 1. D'après la loi des grands nombres, cette proportion fournit une valeur approchée de la probabilité de "tomber" à l'intérieur du disque. On en déduit ainsi (cf question 1) ci-dessous) une valeur approchée de π, on dit une "estimation de π". De plus, le théorème central limite permet de prédire, en fonction du nombre de points lancés, avec une probabilité de 95% l'écart maximal entre la (vraie) valeur de π et la valeur approchée obtenue.

1. Tracer sur une feuille de brouillon "à main levée" le carré $[-1, 1] \times [-1, 1]$ et le disque de centre $(0, 0)$ et de rayon 1.

[3] A titre exceptionnel, on pourra prendre le temps de se renseigner directement sur Internet durant le TP !

Combien vaut l'aire du carré? l'aire du disque?
Quelle est la probabilité qu'un point lancé au hasard et de façon uniforme dans le carré "tombe" à l'intérieur du disque?
Comment traduire qu'un point de coordonnées (x, y) est situé à l'intérieur du disque?

2. Pour une valeur de n à fixer ($n > 35$), simuler le lancer de n points dans le carré (*il suffit de choisir séparément (indépendamment) les abscisses et les ordonnées au hasard dans* $[-1, 1]$).
   Représenter ces n points dans un graphique.
   Calculer la proportion de points tombés à l'intérieur du disque et stocker l'estimation de π ainsi obtenue.
3. Renouveler l'estimation de la question précédente jusqu'à obtenir N estimations de π (prendre par exemple $N = 1000$), autrement dit N valeurs approchées.
   En éliminant les estimations "trop grandes" et "trop petites", dans quel intervalle se situent 95% des estimations?
4. En utilisant le TCL, calculer de façon théorique l'intervalle de fluctuation centré sur π dans lequel doivent se trouver 95% des estimations. Comparer avec la question précédente.

XII Estimation de Taux de Fertilité :

On considère un échantillon de 169 brebis de la race "Ile de France". Ces brebis ont été mises en lutte, on a obtenu 108 brebis pleines (c'est à dire fécondées).
A l'issue de cette expérience, on cherche à estimer le taux de fertilité t de la race "Ile de France".
QCM1 : Soit $\hat{t}$ l'estimateur de t obtenu par la méthode du maximum de vraisemblance. Peut-on affirmer que

A $\hat{t}$ est un estimateur sans biais de t ?

B $\hat{t}$ est un estimateur biaisé de t ?

C $\hat{t} = 69,2\%$?

D $\hat{t} = 63,9\%$?

E $\hat{t} = 59,2\%$?

QCM2 : Un intervalle de confiance au niveau 95% pour ce taux t est donné par

A $I = [0,636; 0,651]$?

B $I = [0,586; 0,651]$?

C $I = [0,566; 0,721]$?

D $I = [0,546; 0,731]$?

E $I = [0,586; 0,601]$?

QCM3 : Peut-on affirmer que

A Un intervalle de confiance au niveau 99% pour t sera plus large que celui obtenu au niveau 95% ?

B Un intervalle de confiance au niveau 90% pour t sera plus large que celui obtenu au niveau 95% ?

C Un intervalle de confiance au niveau 99% pour t sera plus étroit que celui obtenu au niveau 95% ?

D Avec 200 brebis plutôt que 169, même si l'estimée ponctuelle de t devait différer de celle obtenue précédemment, on serait en mesure de fournir un intervalle de confiance plus étroit pour t au niveau 95% ?

E Avec 500 brebis plutôt que 169, si l'estimée ponctuelle de t est proche de celle obtenue précédemment, on sera en mesure de fournir un intervalle de confiance plus étroit pour t au niveau 95% ?

X Taux d'échec d'un vaccin :

Des études antérieures ont révélé que le taux d'échec τ associé à une certaine vaccination est situé entre 10% et 15%.

On prépare une expérience dans le but de déterminer à 1% près la proportion de sujets non-immunisés par ce vaccin, en acceptant un coefficient de risque $\alpha = 0,05$. n personnes seront donc vaccinées, et l'on relèvera par la suite qui parmi ces n personnes n'est pas immunisé à l'issue de la vaccination.

QCM4 : Peut-on affirmer que

A Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit une loi de Poisson ?

B Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit une loi binomiale ?

C Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit approximativement une loi de Poisson ?

D Les variables "nombre de sujets non-immunisés" et "nombre de sujets immunisés" sont indépendantes ?

E Les variables "nombre de sujets non-immunisés" et "nombre de sujets immunisés" sont négativement corrélées ?

QCM5 : Pour parvenir à ses fins (déterminer à 1% près et au niveau de confiance 95% la proportion de sujets non-immunisés), quel est le nombre minimal n_0 de sujets que l'on doit vacciner ?

A $n_0 = 100$

B $n_0 = 1000$

C $n_0 = 3900$

D $n_0 = 4900$

E $n_0 = 5400$

XIII Encore des sondages :

A la veille d'une consultation électorale majeure, on a interrogé une centaine d'électeurs constituant un échantillon représentatif. 58 d'entre eux ont déclaré avoir l'intention de voter pour le candidat *Toulemonde*.

QCM6 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A $I = [0,453; 0,707]$ constitue un intervalle de confiance au niveau 95% pour la cote de M. Toulemonde.

B $I = [0,483; 0,677]$ constitue un intervalle de confiance au niveau 95% pour la cote de M. Toulemonde.

C $I = [0,453; 0,707]$ constitue un intervalle de confiance au niveau 99% pour la cote de M. Toulemonde.

D Au niveau de confiance 99%, on peut affirmer que M. Toulemonde va remporter ces élections.

E Au niveau de confiance 99%, on n'est pas en mesure de garantir que M. Toulemonde va remporter ces élections.

QCM7 : Pour une même fréquence observée d'électeurs favorables à M. Toulemonde, quelle devrait être la taille minimale n_0 de l'échantillon permettant d'affirmer au niveau de confiance 95% que M. Toulemonde sera élu ?

A $n_0 = 100$

B $n_0 = 103$

C $n_0 = 147$

D $n_0 = 200$

E $n_0 = 10000$

XIV Estimations de valeurs moyennes au sein de deux populations :

Dans une région d'Europe, l'étude de la masse du cerveau mesurée en grammes chez des sujets âgés de 20 à 49 ans a conduit aux résultats suivants :

Hommes

Val. approx.	***1170***	***1220***	***1270***	***1320***	***1370***	***1420***	***1470***	**Total**
Effectifs	***5***	***36***	***45***	***50***	***61***	***49***	***19***	***265***

Femmes

Val. approx.	***1070***	***1120***	***1170***	***1220***	***1270***	***1320***	***1370***	**Total**
Effectifs	***12***	***22***	***45***	***54***	***52***	***20***	***10***	***215***

QCM8 : Quel est l'intervalle de confiance au niveau 99% pour la valeur moyenne de cette masse au sein de la population masculine ?

A $I = [1326; 1345]$

B $I = [1323; 1349]$

C $I = [1328; 1344]$

D $I = [1322; 1350]$

E $I = [1320; 1351]$

QCM9 : Quel est l'intervalle de confiance au niveau 99% pour la valeur moyenne de cette masse au sein de la population féminine ?

A $I' = [1203; 1230]$

B $I' = [1209; 1233]$

C $I' = [1209; 1229]$

D $I' = [1206; 1233]$

E $I' = [1206; 1330]$

QCM10 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A Pour l'ensemble de la population, une estimation ponctuelle de cette masse moyenne s'obtient en effectuant la demi-somme $\frac{1}{2}(\hat{m}_h + \hat{m}_f)$.

B Pour l'ensemble de la population, une estimation ponctuelle de cette masse moyenne s'obtient en effectuant une somme $\frac{265}{480}\hat{m}_h + \frac{215}{480}\hat{m}_f$.

C Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est moins large que pour la seule population masculine ?

D Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est plus large que pour la seule population masculine ?

E Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est de même largeur que pour la seule population masculine ?

XV Estimation de la fréquentation d'un hôpital :

On considère un hôpital comportant cent salles de consultation. Chacune de ces salles accueille quotidiennement un nombre de patients qui est modélisé par une variable de Poisson de paramètre $\lambda = 10$, et l'on suppose que ces variables de Poisson sont indépendantes. Soit

S le nombre total de patients qui viennent pour une consultation à l'hôpital un jour donné, puis $M = S/100$ le nombre moyen de consultations par salle ce jour-là.

QCM11 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A $\mathbb{E}(M) = 0,1$

B $\mathbb{E}(M) = 10$

C $\mathbb{E}(S) = 100$

D $\mathbb{E}(S) = 1000$

E $\mathrm{Var}(S) = 10000$

QCM12 : Quelle est la probabilité de voir S dépasser la valeur 1050 un jour donné (à 10^{-2} près) ?

A $0,06$

B $0,11$

C $0,26$

D $0,48$

E $0,96$

XVI Estimation d'un paramètre poissonien :

On suppose que le nombre d'accidents survenant dans une certaine ville un jour donné suit une loi de Poisson de paramètre λ inconnu. Pendant un an, on a relevé quotidiennement le nombre X d'accidents survenus dans cette ville durant la journée, le tableau ci-dessous résume les résultats de cette enquête :

Nombre d'accidents dans la journée

Val. observ.	***0***	***1***	***2***	***3***	***Total***
Nombre de j.	***200***	***100***	***55***	***10***	***365***

On part du principe que les variables $X_1, X_2, \ldots, X_{365}$ associées à chaque journée sont indépendantes, et l'on note M la variable aléatoire de moyenne annuelle des nombres d'accidents quotidiens.

QCM13 : Peut-on affirmer que

A M suit exactement une loi de Poisson ?

B M suit approximativement une loi binomiale ?

C M suit approximativement une loi normale ?

D $\mathbb{E}(M) = \lambda$?

E $\mathrm{Var}(M) = \frac{\lambda}{365}$?

QCM14 : On note s^2 la variance empirique obtenue pour ces variables de Poisson $X_1, X_2, \ldots, X_{365}$ et $\hat{x}$ leur moyenne empirique. Peut-on affirmer que

A $\hat{x} = 0,66$?

B $\hat{x} = 91,3$?

C $s^2 = 0,69$?

D $s^2 = 6,85$?

E $s^2 = 68,5$?

QCM15 : On choisit de noter $I_{1-\alpha}$ un intervalle de confiance de niveau $(1-\alpha)$ pour le paramètre λ. En arrondissant les bornes à 10^{-2} près, peut-on affirmer que

A $I_{0,95} = [0,57; 0,75]$?

B $I_{0,95} = [0,64; 0,68]$?

C $I_{0,95} = [90,5; 92,1]$?

D $I_{0,90} = [0,59; 0,73]$?

E $I_{0,90} = [90,8; 91,8]$?

XVII Mesures et précision :

Un appareil dose la concentration d'une substance sans biais, mais avec une erreur de mesure suivant une loi normale de moyenne nulle et de variance $100\,(mg/L)^2$. Si on effectue plusieurs mesures d'une concentration sur un même prélèvement, on suppose que toutes les erreurs de mesures sont indépendantes. En répétant les mesures puis en choisissant leur moyenne comme résultat, on souhaite obtenir une précision (demi-longueur d'int. de confiance de niveau 95%) inférieure ou égale à $5\,mg/L$.

QCM16 : Combien faut-il effectuer de mesures pour obtenir la précision désirée ?

A 5

B 10

C 12

D 16

E 20

QCM17 : Avec l'appareil de la question précédente, le coût d'une mesure est de 100 Euros. On se demande s'il est rentable d'acheter un nouvel appareil pour doser la concentration de la substance. Un nouvel appareil produit des mesures sans biais et comportant une erreur de variance $40\,(mg/L)^2$ seulement, mais le coût de chaque mesure avec ce nouvel appareil s'élève à 200 Euros. Ici encore, l'objectif est d'obtenir une précision inférieure ou égale à $5\,mg/L$, quitte à combiner de nombreuses mesures. Peut-on affirmer que

A Pour atteindre ce but, il faut effectuer 5 mesures avec le nouvel appareil ?

B Pour atteindre ce but, il faut effectuer 7 mesures avec le nouvel appareil ?

C Pour atteindre ce but, il faut effectuer 9 mesures avec le nouvel appareil ?

D Il vaut mieux garder l'ancien appareil de dosage.

E Il vaut mieux acheter le nouvel appareil de dosage.

XVIII Encore des sondages :

Une élection oppose deux candidats, et tout porte à croire qu'elle sera très serrée : on part du principe que tous les électeurs vont voter pour l'un ou l'autre candidat, et que les scores de chacun d'eux seront proches de 50%. On souhaite effectuer un sondage à la sortie des bureaux de vote.

QCM18 : Combien d'électeurs faut-il interroger pour atteindre une précision (demi-longueur d'intervalle de confiance au niveau 95%) de $0,02$?

A 996

B 1010

C 1563

D 2401

E 2542

QCM19 : Si l'on choisit d'interroger 1000 électeurs, quelle sera la précision obtenue pour les proportions de voix remportées par chaque candidat ?

A $0,01$

B $0,03$

C $0,05$

D $0,07$

E $0,10$

6 STATISTIQUE INFERENTIELLE (II) : TESTS D'HYPOTHESES

Comme nous allons le voir dans ce dernier chapitre, les Tests Statistiques s'utilisent très couramment dans les sciences expérimentales et leur mise en oeuvre requiert d'apprendre à appliquer quelques méthodes de calcul plutôt simples.

6.1 Introduction : des Intervalles de Confiance aux Tests

Le but poursuivi est de parvenir à "trancher" entre deux hypothèses contradictoires (H_0) et (H_1) au vu des valeurs $x_1, x_2, \ldots, x_n$ prises par certaines variables dans une expérience, et le principe suivi sera toujours le même :

- à partir de ces valeurs, évaluer une variable de test $V = V(x_1, x_2, \ldots, x_n)$.
- si la variable de test $V = V(x_1, x_2, \ldots, x_n)$ prend une valeur "anormale" (anormalement basse ou anormalement élevée) au regard de l'hypothèse (H_0), on rejette cette hypothèse (H_0) (et on accepte l'hypothèse (H_1)).
- si, en revanche, la variable de test V prend une valeur "raisonnable" (ni trop basse ni trop élevée) au regard de cette hypothèse (H_0), on ne peut pas rejeter (H_0).

Pour être plus complet, il convient de considérer que l'on est dans une situation d'***information incomplète*** : même avec un très vaste échantillon, on ne peut pas être totalement certain de la validité de (H_0) ou de celle de (H_1), une part d'incertitude demeure.
Il y a donc dans cette situation deux manières de se tromper : on pourra *rejeter l'hypothèse* (H_0) *alors que celle-ci est vraie* (***erreur de première espèce***) ou encore *accepter* (H_0) *alors que celle-ci est fausse* (***erreur de seconde espèce***). D'où les définitions suivantes :

- Le ***risque de 1ère espèce*** α (ou risque) correspond à la probabilité sous (H_0) de rejeter (H_0).
- Le ***risque de seconde espèce*** β correspond à la probabilité sous (H_1) de rejeter (H_1).
- La ***puissance*** du test vaut par définition $(1 - \beta)$.

Dans la pratique, l'hypothèse (H_0) est une "hypothèse de référence" sous laquelle les calculs de probabilités peuvent être menés de façon satisfaisante, on cherchera donc en priorité à faire en sorte que le risque de 1ère espèce α ne soit pas trop grand.
En général, un *seuil de risque* (e.g. 5% ou 1%) est prescrit, le but est d'alors d'utiliser un test ayant un risque α majoré par ce seuil.
Venons-en tout de suite à un premier exemple inspiré du chapitre précédent.

Exemple 1 : Sondages d'opinion
Les candidats BO et NG doivent s'affronter lors des prochaines élections, un institut de sondage a demandé à 1000 personnes pour qui elles s'apprêtaient à voter ; 450 personnes ont répondu en faveur de NG, et les 550 autres en faveur de BO.
Dans ces conditions, peut-on accepter l'hypothèse
(H_0) : ***"BO va remporter les élections"*** ?

D'un point de vue probabiliste, le score réalisé par BO sur un échantillon de 1000 électeurs suit une loi $\mathcal{B}(n = 1000; p)$; le second paramètre p de cette loi binomiale nous est inconnu mais nous pourrions, à partir de l'estimée ponctuelle $\hat{p} = \frac{550}{1000} = 0,55$, fabriquer un intervalle de confiance au niveau 5% ou 1% centré en $\hat{p}$.
Au lieu de s'intéresser à un tel intervalle, on considère plutôt un intervalle de valeurs "anormales" sous l'hypothèse (H_0) : sous cette hypothèse, il serait franchement étonnant que la variable de test $V = \hat{p}$ prenne une valeur "plutôt basse".
Il nous faut donc construire un ***"intervalle de rejet"*** $I_R = [0; a]$, ou ***"zone de rejet"***, que l'on utilisera de la manière suivante :

- si $V \in I_R$, autrement dit si $\hat{p} \leq a$, on rejette l'hypothèse (H_0).
- en revanche, si $\hat{p} > a$, on accepte (H_0).

La valeur du seuil de rejet a se règle en fait à partir du seuil de risque que l'on est prêt à accepter.
Par exemple, si notre but est d'avoir $\alpha \leq 0,05$, en posant

$$\alpha = \mathbb{P}_{H_0}\{V \in I_R\} = \mathbb{P}_{H_0}\{\overline{X}_n \leq a\},$$

on a

$$\alpha \leq \mathbb{P}_{p=0,5}\{\overline{X}_n \leq a\} = \mathbb{P}_{p=0,5}\{\frac{\overline{X}_n - 0,5}{\sqrt{\frac{0,5\times 0,5}{10^3}}} \leq \frac{a - 0,5}{\sqrt{\frac{0,5\times 0,5}{10^3}}}\} \simeq \mathbb{P}_{p=0,5}\{Z \leq \frac{a - 0,5}{\sqrt{\frac{0,5\times 0,5}{10^3}}}\},$$

où Z est une variable normale standard (cf. TCL!).
Notre but est donc de faire en sorte que

$$\frac{a - 0,5}{\sqrt{\frac{0,5\times 0,5}{10^3}}} = 1,645$$

ce qui nous donne

$$a = 0,5 + \frac{1,645}{2 \times 10^{3/2}} = 0,526$$

Puisque le score observé pour BO dépasse le seuil de $52,6\%$ (ce score sélève à 55%), on est en mesure d'accepter l'hypothèse (H_0) au risque 5%.

Voici un second exemple plus proche de nos préoccupations médicales et pharmacologiques.

Exemple 2 : Procès suite à une campagne de vaccination

Des plaignants ont poursuivi en justice le ministère israélien de la Santé suite à une campagne de vaccination menée sur des enfants et ayant entraîné des dommages fonctionnels irréversibles pour certains d'entre eux [4].

Ce vaccin était connu pour entraîner ce type de dommages en de très rares circonstances. Des études antérieures menées dans d'autres pays semblaient indiquer que ce risque était d'un cas sur $310'000$ vaccinations. Les plaignants avaient été informés de ce risque et l'avaient accepté.

Les doses de vaccin ayant provoqué les dommages faisant l'objet de la plainte provenaient d'un lot qui avait servi à la vaccination de $300'533$ enfants ; dans ce groupe, quatre cas de dommages ont été détectés suite à la vaccination.

Convient-il d'***accepter l'hypothèse*** (H_0) ***selon laquelle ces données sont conformes au risque présenté par les autorités sanitaires*** ? Ou bien faut-il plutôt ***opter pour l'hypothèse selon laquelle le nombre d'enfants victimes de dommages est bien trop élevé pour cela*** ?

Observons tout d'abord que le comptage du nombre d'enfants subissant des dommages fonctionnels suite à la vaccination peut être modélisé mathématiquement par une variable de Poisson de paramètre $\lambda = n \cdot p$, où $n = 300'533$ est la taille du groupe d'enfants vaccinés, tandis que p est la probabilité pour un enfant de subir de tels dommages.

Traduite mathématiquement, l'hypothèse (H_0) sécrit donc : $\lambda = np = np_0 = \frac{300533}{310'000} = 0,969$ (valeur théorique annoncée par les autorités sanitaires), et l'hypothèse contradictoire (H_1) qui nous intéresse s'écrit quant à elle : $\lambda > np_0$.

La "zone de rejet" (ou intervalle de rejet) correspondant au rejet de l'hypothèse (H_0) est alors un intervalle de la forme $I_R = [N; +\infty[$, où le seuil N est à régler en fonction du risque de première espèce de notre test. L'idée étant simplement de rejeter (H_0) si l'on observe un trop grand nombre d'enfants victime de ces dommages, comment régler le seuil N pour obtenir un risque (de première espèce) α majoré par 5%, par exemple ?

Il s'agit en fait de fixer N en sorte que $\mathbb{P}_{H_0}\{nb.\ dommages\ apparus \geq N\} \leq 0,05$; or, avec $\lambda = 0,969$, on a

$$\mathbb{P}_{H_0}(nb.\ dommages\ apparus \geq N) = 1 - \sum_{k=0}^{N-1} e^{-\lambda}\frac{\lambda^k}{k!} = \begin{cases} 0,0747 & \text{si } N = 3 \\ 0,0171 & \text{si } N = 4 \end{cases}$$

Pour que le risque α soit majoré par 5%, il convient donc de fixer $N = 4$; la zone de rejet est donc l'intervalle $I_R = [4; +\infty[$, et puisque l'on a enregistré quatre cas d'apparition de dommages fonctionnels, l'hypothèse (H_0) doit être rejetée au seuil de risque $\alpha = 5\%$.

[4] cf Murray Aitkin, *Evidence and the Posterior Bayes Factor*, Math. Scientist (1992)

Récapitulatif : retenons les principes suivants :

1. Les ***Tests Statistiques*** sont utilisés dans le but de savoir s'il convient d'***accepter une hypothèse de référence*** (H_0) ou alors de rejeter cette hypothèse pour accepter une hypothèse (H_1) contredisant (H_0).
2. Cette acceptation ou rejet de l'hypothèse (H_0) s'opère après avoir calculé la valeur prise par une ***variable de test*** $V = V(x_1, x_2, \ldots, x_n)$ en fonction des observations $x_1, x_2, \ldots, x_n$: si la valeur prise par V est située en dehors d'une certaine ***"Zone de Rejet"*** I_R, il convient d'accepter (H_0), tandis que si cette valeur est située dans I_R on rejette (H_0) pour accepter plutôt l'hypothèse (H_1).
3. En fait, il s'agit d'accepter ou rejeter (H_0) pour un certain ***seuil de risque*** (de 1ère espèce) α : l'intervalle I_R est défini en sorte que

$$\mathbb{P}_{H_0}(V \in I_R) \leq \alpha$$

La définition d'une zone de rejet I_R requiert en général d'utiliser un "Théorème Limite" du Calcul des Probabilités. Nous avons vu un tel théorème dans ce cours, le TCL, où les variables normales jouent un rôle primordial. Nous allons donc commencer par considérer des Tests Statistiques construits à partir du TCL.

6.2 Tests utilisant la loi $\mathcal{N}(0;1)$

Il y a pour l'essentiel quatre types de tests reposant sur le TCL et tels que, sous l'hypothèse (H_0), la variable de test V suit une loi normale standard.

6.2.1 Tests de conformité à une fréquence théorique

I- Situation expérimentale et hypothèse de référence :
dans une population de n individus, on a relevé la réalisation ou l'absence de réalisation d'un certain événement E en posant $x_i = 1$ si cet événement est réalisé pour l'individu i et $x_i = 0$ sinon.
L'hypothèse de référence (H_0) est que la fréquence observée $\overline{x}_n$ pour les réalisations de l'événement E est conforme à une certaine fréquence théorique p_0.

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire $V = \frac{\overline{X}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ suit approximativement la loi d'une variable $\mathcal{N}(0;1)$ Z, tout du moins si n est "suffisamment grand" ($n \geq 30$, avec $np_0 \geq 5$ et $n(1-p_0) \geq 5$).

III- Valeur à calculer :
il convient donc de calculer $V = V(x_1, x_2, \ldots, x_n) = \frac{\overline{x}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, puis de rejeter l'hypothèse de référence (H_0) si la valeur V obtenue est "anormalement élevée" ou encore "anormalement basse".

IV- Définition de la Zone de Rejet :
1. Dans le cas (le plus courant) d'un test bilatéral :
l'hypothèse alternative (H_1) est $p \neq p_0$, où p désigne la prob. de réalisation de l'événement E. On est alors conduit à construire une zone de rejet I_R du type : $I_R = \mathbb{R}\backslash]-\varepsilon_\alpha; \varepsilon_\alpha[$, où le quantile ε_α est choisi en sorte que $\mathbb{P}\{|Z| > \varepsilon_\alpha\} = \alpha$.
2. Dans le cas d'un test unilatéral :
si l'on sait a priori que cette probabilité p doit vérifier $p \geq p_0$, l'hypothèse alternative (H_1) devient $p > p_0$. On est alors conduit à construire une zone de rejet I_R du type : $I_R =]x_\alpha; +\infty[$, le quantile $x_\alpha = \varepsilon_{2\alpha}$ étant choisi en sorte que $\mathbb{P}\{Z > x_\alpha\} = \alpha$.

Voyons tout de suite deux exemples d'application de cette méthode :

Exemple 3 : fréquences de rhésus chez certaines populations
Dans la population française, le pourcentage d'individus dont le sang est de rhésus négatif s'élève à 15%.
Dans un échantillon représentatif de 200 Basques français, on observe que 44 individus sont

de rhésus négatif.
Au vu de ces résultats, peut-on affirmer au risque $\alpha = 5\%$ que les Basques diffèrent du reste de la France pour ce qui est de la variable "rhésus" ?

Il s'agit d'appliquer un test de conformité entre une fréquence observée $\overline{x}_n = \frac{44}{200} = 0,22$ et la probabilité théorique $p_0 = 0,15$.
L'hypothèse de référence (H_0) s'écrit donc mathématiquement : $p = p_0 = 0,15$. Accepter (H_0) revient ici à considérer que la différence observée entre $\overline{x}_n$ et p_0 s'explique raisonnablement en considérant les fluctuations dues au hasard.
Puisqu'il n'y a pas de raison de considérer a priori que la probabilité p doit être au moins égale à p_0, ou encore au plus égale à p_0, on applique un test de conformité bilatéral : l'hypothèse alternative (H_1) est donnée simplement par $p \neq p_0$ et on pourra se permettre de rejeter (H_0) au risque 5% si $V = V(x_1, x_2, \ldots, x_n) = \frac{\overline{x}_n - 0,15}{\sqrt{\frac{0,15(0,85)}{200}}}$ figure dans la zone de rejet $I_R = \mathbb{R}\backslash]-\varepsilon_{0,05}; \varepsilon_{0,05}[$, c'est à dire si $|V(x_1, x_2, \ldots, x_n)| \geq \varepsilon_{0,05} = 1,96$.
Calculons la valeur de V : on a $V = \frac{0,22-0,15}{\sqrt{\frac{0,15(0,85)}{200}}} \approx 2,77 > 1,96$.
Au risque 5%, il convient donc de rejeter (H_0) et de conclure que les Basques diffèrent du reste de la France en ce qui regarde la variable "rhésus".

Exemple 4 : efficacité d'une crème anti-rides
On considère une population où le pourcentage d'individus présentant des rides s'élève à 25%, et l'on demande à 200 individus choisis au hasard d'utiliser une crème anti-rides. Quelque temps plus tard, on constate que 40 personnes parmi ces 200 individus présentent des rides.
Peut-on affirmer au risque $\alpha = 5\%$ que ce traitement anti-rides est sans effet ?

Il s'agit à nouveau d'appliquer un test de conformité entre une fréquence observée et une fréquence théorique, mais cette fois-ci l'hypothèse alternative s'écrit $p < p_0$ (*"Ce traitement produit un effet"*) et le test de conformité est unilatéral : il convient de rejeter l'hypothèse (H_0) si la variable V prend une valeur trop basse, plus précisément si $V \in I_R =]-\infty; -1,645[$ (puisque $\mathbb{P}\{Z \in]-\infty; -1,645[\} = \mathbb{P}\{Z < -1,645\} = 0,05$).
Or le calcul de V donne ici $V = \frac{\frac{40}{200} - 0,25}{\sqrt{\frac{0,25(0,75)}{200}}} \approx -1,63 \geq -1,645$. Donc, au risque $\alpha = 5\%$, on n'est pas en mesure de rejeter l'hypothèse (H_0) d'inefficacité de ce traitement.

6.2.2 Tests d'égalité entre deux fréquences

I- Situation expérimentale et hypothèse de référence :
on relève la réalisation ou l'absence de réalisation d'un certain événement E au sein de deux populations de tailles respectives n_1 et n_2, en posant $x_i = 1$ si cet événement est réalisé pour l'individu i de la 1ère population et $x_i = 0$ sinon, et $y_j = 1$ si cet événement

est réalisé pour l'individu j de la seconde population, tandis que $y_j = 0$ sinon.
L'hypothèse de référence (H_0) est que la probabilité de réalisation de E est identique au sein des deux populations : $p_1 = p_2 = p_0$.

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire $V = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}}}$ suit approximativement la loi d'une variable $\mathcal{N}(0;1)$ Z, tout du moins si n_1 et n_2 sont "suffisamment grands".

III- Valeur à calculer :
il convient donc de calculer

$$V = V(x_1, x_2, \ldots, x_{n_1}; y_1, y_2, \ldots, y_{n_2}) = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}},$$

où $\hat{p} = \frac{n_1}{n_1+n_2}\overline{x}_{n_1} + \frac{n_2}{n_1+n_2}\overline{y}_{n_2} = \frac{1}{n_1+n_2}\left(\sum_i x_i + \sum_j y_j\right)$ désigne l'estimation de p_0 par la méthode du max. de vraisemblance. On pourra alors rejeter l'hypothèse de référence (H_0) si la valeur V obtenue est "anormalement élevée" ou encore "anormalement basse".

IV- Définition de la Zone de Rejet :
1. Dans le cas (le plus courant) d'un test bilatéral :
l'hypothèse alternative (H_1) est $p_1 \neq p_2$, et on est conduit à construire une zone de rejet I_R du type : $I_R = \mathbb{R}\backslash]-\varepsilon_\alpha; \varepsilon_\alpha[$, où le quantile ε_α est choisi en sorte que $\mathbb{P}\{|Z| > \varepsilon_\alpha\} = \alpha$.
2. Dans le cas d'un test unilatéral :
si l'on sait a priori que les probabilités p_1 et p_2 doivent vérifier $p_1 \geq p_2$, l'hypothèse alternative (H_1) devient $p_1 > p_2$. On est alors conduit à construire une zone de rejet I_R du type : $I_R =]x_\alpha; +\infty[$, le quantile $x_\alpha = \varepsilon_{2\alpha}$ étant choisi en sorte que $\mathbb{P}\{Z > x_\alpha\} = \alpha$.

Voici deux exemples d'application de cette méthode :

Exemple 5 : comparaison d'efficacités pour un traitement
A la suite d'un même traitement, on a observé 40 bons résultats chez 70 malades jeunes et 50 bons résultats chez 100 malades âgés.
Au risque $\alpha = 10\%$, peut-on affirmer qu'il n'y a pas de lien entre l'âge du malade et l'efficacité du traitement ?

Il s'agit d'utiliser les fréquences observées $\overline{x} = \frac{40}{70}$ et $\overline{y} = \frac{50}{100}$ pour tester l'hypothèse nulle $p_1 = p_2$.
Comme il n'y a pas de raison a priori d'estimer que le traitement fonctionne au moins aussi bien au sein d'une population qu'au sein de l'autre, on choisit ici un test bilatéral.
L'estimation de $p_1 = p_2$ par la méthode du max. de vraisemblance donne ici $\hat{p} = \frac{40+50}{70+100} =$

$\frac{9}{17}$, on obtient donc ensuite

$$V = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{\frac{4}{7} - 0,5}{\sqrt{\frac{9}{17}\frac{8}{17}(\frac{1}{70} + \frac{1}{100})}} \approx 0,92$$

Pour $\alpha = 10\%$, on lit $\varepsilon_\alpha = 1,645$. Puisque $V \in [-\varepsilon_\alpha; \varepsilon_\alpha]$, on peut accepter l'hypothèse (H_0) au risque 10%.

Exemple 6 : efficacité d'un dispositif de protection
Dans des services de maladies infectieuses, on observe des contaminations parmi les 2'100 employés constituant le personnel infirmier. On impose à 50 de ces personnes, désignées au hasard, des mesures de protection particulières ; on observe par la suite 7 contaminations chez ces 50 personnes.
Par ailleurs, parmi 50 autres personnes non-protégées et désignées au hasard, on a observé 11 contaminations durant la même période.
A quel risque α peut-on conclure à l'efficacité du dispositif de traitement ?

Il s'agit de comparer deux fréquences expérimentales $\overline{x}$ et $\overline{y}$, où $\overline{x} = \overline{x}_{n_1} = 0,14$ et $\overline{y} = \overline{y}_{n_2} = 0,22$ avec $n_1 = n_2 = 50$. Mais ici la situation est unilatérale a priori : on va tester l'hypothèse nulle $p_1 = p_2$ contre l'hypothèse alternative $p_1 < p_2$, et l'on cherche la plus petite valeur du risque (de 1ère espèce) α permettant de rejeter (H_0) au bénéfice de (H_1).
Les calculs donnent ici

$$V = \frac{\overline{x} - \overline{y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0,14 - 0,22}{\sqrt{0,18(1-0,18)(\frac{1}{50} + \frac{1}{50})}} \approx -1,041,$$

et d'après nos tables, pour une variable normale standard Z : $\mathbb{P}\{Z < -1,041\} \simeq 0,15$.
Il n'est donc pas possible de rejeter l'hypothèse (H_0) pour conclure à l'efficacité de ce dispositif autrement qu'en acceptant un risque de 15%. Il s'agit là d'un risque élevé, on peut donc estimer que les données statistiques obtenues ne permettent pas de conclure à l'efficacité du dispositif de protection.

6.2.3 Tests de conformité à une moyenne théorique

I- Situation expérimentale et hypothèse de référence :
On relève les valeurs $x_1, x_2, \ldots, x_n$ prises par des v.a. i.i.d. $X_1, X_2, \ldots, X_n$ telles que $\mathbb{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = \sigma^2$. Ces observations permettent de calculer la moyenne empirique $\overline{x}_n$ puis la variance empirique

$$s^2 = s(x_1, x_2, \ldots, x_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\cdot\overline{x}_n^2$$

Mathématiquement, l'hypothèse de référence (H_0) s'écrit $\mu = \mu_0$, où μ_0 est une valeur connue.

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire $V = \frac{\overline{X}_n - \mu_0}{(\sigma/\sqrt{n})}$ suit approximativement la loi d'une variable $\mathcal{N}(0;1)$ Z, tout du moins si n est "suffisamment grand" $(n \geq 30)$.

III- Valeur à calculer :
il convient donc de calculer $V = V(x_1, x_2, \ldots, x_n) = \frac{\overline{x}_n - \mu_0}{(s/\sqrt{n})}$, puis de rejeter l'hypothèse de référence (H_0) si la valeur V obtenue est "anormalement élevée" ou encore "anormalement basse".

IV- Définition de la Zone de Rejet :
1. Dans le cas d'un test bilatéral :
l'hypothèse alternative (H_1) est $\mu \neq \mu_0$, on est donc conduit à construire une zone de rejet I_R du type : $I_R = \mathbb{R}\backslash]-\varepsilon_\alpha; \varepsilon_\alpha[$, où le quantile ε_α est choisi en sorte que $\mathbb{P}\{|Z| > \varepsilon_\alpha\} = \alpha$.
2. Dans le cas d'un test unilatéral :
si l'on sait a priori que le paramètre μ doit vérifier $\mu \geq \mu_0$, l'hypothèse alternative (H_1) devient $\mu > \mu_0$. On est alors conduit à construire une zone de rejet I_R du type : $I_R =]x_\alpha; +\infty[$, le quantile $x_\alpha = \varepsilon_{2\alpha}$ étant choisi en sorte que $\mathbb{P}\{Z > x_\alpha\} = \alpha$.

Voyons maintenant un exemple d'application de cette méthode :

Exemple 7 : conformité d'un dosage de substance active
Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir $2,5\,g$ de substance active.
Cent comprimés sont choisis au hasard dans une certaine production puis analysés ; il s'avère que ces comprimés contiennent en moyenne $2,6\,g$ de substance active, avec un écart-type empirique s valant $0,4\,g$.
Au risque $\alpha = 5\%$, peut-on considérer que cette production respecte les spécifications prescrites ?

Il s'agit de tester la conformité d'une moyenne empirique $\overline{x} = 2,6$ avec une moyenne théorique μ_0, et ce test est bilatéral (la présence d'un excédent de substance active n'est pas considéré comme bénéfique ou même acceptable a priori). Les calculs donnent dans le cas présent :

$$V = \overline{x}_n - \mu_0 s/\sqrt{n} = 2,6 - 2,50,4/\sqrt{100} = 2,5$$

Puisque $\varepsilon_{0,05} = 1,96$, on peut rejeter (H_0) et considérer, au risque 5%, que cette production ne respecte pas les spécifications.

6.2.4 Tests d'égalité entre deux moyennes

I- Situation expérimentale et hypothèse de référence :
On relève les valeurs $x_1, x_2, \ldots, x_{n_1}$ prises par des v.a. i.i.d. $X_1, X_2, \ldots, X_{n_1}$ telles que $\mathbb{E}(X_i) = \mu_1$ et $\text{Var}(X_i) = \sigma_1^2$, mais aussi les valeurs $y_1, y_2, \ldots, y_{n_2}$ prises par des v.a. i.i.d. $Y_1, Y_2, \ldots, Y_{n_2}$ telles que $\mathbb{E}(Y_j) = \mu_2$ et $\text{Var}(Y_j) = \sigma_2^2$. On suppose en outre qu'il n'y a pas de dépendance entre les variables du premier et du deuxième type.
Mathématiquement, l'hypothèse de référence (H_0) s'écrit $\mu_1 = \mu_2$.

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire $V = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ suit approximativement la loi d'une variable $\mathcal{N}(0;1)$ Z, tout du moins si n_1 et n_2 sont "suffisamment grands".

III- Valeur à calculer :
il nous faut donc calculer $V = \dfrac{\overline{x}_{n_1} - \overline{y}_{n_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, puis rejeter l'hypothèse de référence (H_0) si la valeur V obtenue est "anormalement élevée" ou encore "anormalement basse".

IV- Définition de la Zone de Rejet :
1. Dans le cas d'un test bilatéral :
l'hypothèse alternative (H_1) est $\mu_1 \neq \mu_2$, on est donc conduit à construire une zone de rejet I_R du type : $I_R = \mathbb{R}\backslash]-\varepsilon_\alpha; \varepsilon_\alpha[$, où le quantile ε_α est choisi en sorte que $\mathbb{P}\{|Z| > \varepsilon_\alpha\} = \alpha$.
2. Dans le cas d'un test unilatéral :
si l'on sait a priori que les paramètres μ_1 et μ_2 doivent vérifier $\mu_1 \geq \mu_2$, l'hypothèse alternative (H_1) devient $\mu_1 > \mu_2$. On est alors conduit à construire une zone de rejet I_R du type : $I_R =]x_\alpha; +\infty[$, le quantile $x_\alpha = \varepsilon_{2\alpha}$ étant choisi en sorte que $\mathbb{P}\{Z > x_\alpha\} = \alpha$.

Voyons enfin un exemple d'application de cette dernière méthode :

Exemple 8 : poids moyens de fruits récoltés à deux époques différentes
On cherche à savoir si les poids de pommes récoltées en début et en fin de saison sont significativement différents. Dans un échantillon de $n_1 = 100$ pommes récoltées en début de saison, on a obtenu un poids moyen de $120\,g$ pour un écart-type empirique s'élevant à $20\,g$. En revanche, pour un second échantillon de $n_2 = 150$ pommes récoltées en fin de saison, le poids moyen obtenu est de $150\,g$ tandis que l'écart-type empirique s'éleve à $10\,g$. Dans ces conditions, doit-on considérer qu'il y a une différence de poids significative entre les fruits de la première et de la seconde récolte ?

Il s'agit d'appliquer un test d'égalité de moyennes, et comme nous n'avons pas de raison d'estimer a priori que les pommes de la seconde récolte sont plus lourdes ou encore plus légères que les autres, ce test est bilatéral.

Le calcul donne ici

$$V = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{120 - 150}{\sqrt{\frac{20^2}{100} + \frac{10^2}{150}}} \simeq -13,88$$

Puisque $\varepsilon_{0,05} = 1,96$, $\varepsilon_{0,01} = 2,576$ et $\varepsilon_{0,001} = 3,291$, on obtient $|V| < u_\alpha$ aussi bien pour un risque α de 5% que pour $\alpha = 1\%$ ou même $\alpha = 0,1\%$!
Il y a donc tout lieu de rejeter l'hypothèse (H_0) d'égalité des poids moyens de la première et de la deuxième récolte.

6.3 Tests utilisant une loi du χ^2

Commençons par définir un nouveau type de v.a. continues : on dira que la variable X suit une *loi du* χ^2 (ou *loi de Pearson*) à ν degrés de liberté si X peut être présentée sous la forme

$$X = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2,$$

pour une suite finie $Z_1, Z_2, \ldots, Z_\nu$ de variables $\mathcal{N}(0;1)$ indépendantes.

Comme on le comprendra au fil de ce dernier paragraphe, les variables de Pearson ont un rôle particulièrement important à jouer en statistique, puisqu'elles apparaissent dans la conclusion de plusieurs théorèmes limites du calcul des probabilités.

6.3.1 Tests de conformité à une loi théorique

Pour ce premier type de tests, le but est d'examiner s'il est raisonnable de modéliser le comportement de certaines grandeurs expérimentales en utilisant des variables aléatoires suivant une loi connue.

I- Situation expérimentale et hypothèse de référence :
Etant donné un système complet d'événements $A_1, A_2, \ldots, A_k$, on relève, pour chacun des n individus d'une population donnée, l'événement réalisé, puis on compte les effectifs correspondant à chacun de ces événements au sein de la population : l'événement A_1 est réalisé par $n_1^{obs.} = Eff._1^{obs.}$ individus, l'événement A_2 par $n_2^{obs.} = Eff._2^{obs.}$ individus, ... , l'événement A_K par $n_k^{obs.} = Eff._k^{obs.}$ individus.

L'hypothèse (H_0) affirme que ces observations sont conformes à une certaine loi mathématique pour laquelle les événements $A_1, A_2, \ldots, A_k$ sont réalisés avec les probabilités $p_1, p_2, \ldots, p_k$. [5] Ces probabilités permettent à leur tour de calculer des "effectifs théoriques" pour chacun des événements $A_1, A_2, \ldots, A_k$ du système complet :
$n_1^{th.} = Eff._1^{th.} = n \cdot p_1$, $n_2^{th.} = Eff._2^{th.} = n \cdot p_2$, ... , $n_k^{th.} = Eff._k^{th.} = n \cdot p_k$.

On dispose donc d'un premier tableau d'*"effectifs observés"* et d'un second tableau d'*"effectifs théoriques"* : il s'agit de comparer ces deux tableaux en évaluant une certaine distance les séparant, puis d'accepter l'hypothèse (H_0) si ces deux tableaux sont "raisonnablement proches" (faible distance) ou de rejeter l'hypothèse (H_0) si ces deux tableaux sont "plutôt éloignés" (distance élevée). Ici encore, le seuil de distance sous lequel (H_0) est acceptée et au-delà duquel (H_0) est rejetée se règle en fonction d'un risque de première espèce que l'on est prêt à accepter.

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k} \frac{(Eff._i^{obs.} - Eff._i^{th.})^2}{Eff._i^{th.}}$$

suit approximativement une loi du χ^2 à $\nu = k - 1 - r$ degrés de liberté, où r désigne le nombre de paramètres estimés pour calculer les probabilités théoriques $p_1, p_2, \ldots, p_k$.

Attention cependant : cette approximation n'a de valeur que dans la mesure où la taille n de la population est raisonnablement grande ($n \geq 30$) ; en outre, on exigera des événements $A_1, A_2, \ldots, A_k$ qu'ils soient suffisamment importants pour que les effectifs théoriques $Eff._1^{th.}, Eff._2^{th.}, \ldots, Eff._k^{th.}$ soient tous ≥ 5. Si tel n'est pas le cas, il faut reprendre ce calcul de distance en regroupant les événements de moindre importance, dans le but d'obtenir un système complet dans lequel tous les effectifs théoriques sont ≥ 5.

[5] Puisque $A_1, A_2, \ldots, A_k$ constitue un système complet d'événements, on a $\sum_{i=1}^{k} p_i = 1$.

III- Valeur à calculer :
il nous faut donc calculer la *"Distance du χ^2"*

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k} \frac{(Eff._i^{obs.} - Eff._i^{th.})^2}{Eff._i^{th.}}$$

puis rejeter l'hypothèse de référence (H_0) si la distance obtenue est "anormalement élevée". En définissant les fréquences observées $f_1^{obs.}, f_2^{obs.}, \ldots, f_k^{obs.}$ par

$$f_1^{obs.} = \frac{n_1^{obs.}}{n} = \frac{Eff._1^{obs.}}{n},\ f_2^{obs.} = \frac{n_2^{obs.}}{n} = \frac{Eff._2^{obs.}}{n},\ \ldots,\ f_k^{obs.} = \frac{n_k^{obs.}}{n} = \frac{Eff._k^{obs.}}{n},$$

on peut tout aussi bien calculer cette distance du χ^2 en termes de fréquences observées et de "fréquences théoriques" (les probabilités $p_1, p_2, \ldots, p_k$), puisque

$$\begin{array}{rcl} V & = & D_{\chi^2}(Tab.Obs.; Tab.Th.) \\ & = & \sum_{i=1}^{k} \frac{(Eff._i^{obs.} - Eff._i^{th.})^2}{Eff._i^{th.}} \\ & = & n \cdot \sum_{i=1}^{k} \frac{(f_i^{obs.} - f_i^{th.})^2}{f_i^{th.}} \\ & = & n \cdot \sum_{i=1}^{k} \frac{(f_i^{obs.} - p_i)^2}{p_i} \end{array}$$

IV- Définition de la Zone de Rejet :
une fois cette distance calculée, il reste à aller rechercher la valeur du quantile q_α associé à une variable X suivant une loi du χ^2 à ν degrés de liberté : le quantile q_α étant tel que $\mathbb{P}\{X > q_\alpha\} = \alpha$, on rejette l'hypothèse nulle de conformité si la distance obtenue est supérieure à q_α, et on accepte cette hypothèse (H_0) sinon.

Voyons un exemple d'application de ces tests de conformité :

Exemple 9 : test d'une hypothèse de répartition poissonienne
Une enquête effectuée auprès de 150 pharmacies a permis d'étudier la fréquentation de ces différents établissements ; durant une période de 15 minutes, on a relevé le nombre d'arrivées de nouveaux clients, pharmacie par pharmacie, pour ensuite dresser le tableau suivant :

Nombre de clients	***0***	***1***	***2***	***3***	***4***	***5***	***6***
Nombre de pharma.	***37***	***46***	***39***	***19***	***5***	***3***	***1***

Au risque 5%, peut-on admettre que le nombre de clients fréquentant l'une de ces pharmacies suit une loi de Poisson ?

X désignant le nombre de clients fréquentant un établissement choisi au hasard, l'hypothèse (H_0) s'écrit aussi : *X suit une loi de Poisson.*
En appelant λ le paramètre de cette loi de Poisson, on a $\mathbb{E}(X) = \mathrm{Var}(X) = \lambda$, et ce paramètre peut être estimé sans biais par la méthode du max. de vraisemblance ; on a

$$\hat{\lambda} = \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{150}\left[37 \times 0 + 46 \times 1 + 39 \times 2 + \ldots + 1 \times 6\right] = 1,48$$

Remarquons qu'ici l'estimation sans biais de la variance donne le résultat suivant :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \approx 1,58,$$

valeur estimée proche de $\hat{\lambda}$; l'hypothèse poissonienne paraît donc tout à fait envisageable.
Dressons donc un tableau des effectifs observés et des effectifs théoriques pour $\lambda = 1,5$:

Evénements	***X=0***	***X=1***	***X=2***	***X=3***	***X=4***	***X=5***	***X>5***
Eff. Obs.	***37***	***46***	***39***	***19***	***5***	***3***	***1***
Probabilités	***0,2231***	***0,3347***	***0,2510***	***0,1255***	***0,0471***	***0,0141***	***0,0045***
Eff. Théoriques	***33,47***	***50,20***	***37,65***	***18,83***	***7,06***	***2,12***	***0,67***

Les effectifs théoriques sont-ils suffisamment proches des effectifs observés pour accepter l'hypothèse (H_0) au risque $\alpha = 5\%$?
Avant de passer au calcul d'une distance du χ^2, remarquons que les deux derniers effectifs théoriques, correspondant aux événements $\{X = 5\}$ et $\{X \geq 6\}$, sont inférieurs à 5 ; il convient donc de regrouper les trois derniers événements ci-dessus en un seul événement $(\{X \geq 4\})$, d'effectif observé 9 et ayant pour effectif théorique $150 \times \mathbb{P}\{X \geq 4\} = 9,85$.
On obtient donc la mesure de distance

$$V = \sum_{i=1}^{k} \frac{(Eff._i^{obs.} - Eff._i^{th.})^2}{Eff._i^{th.}} = \frac{(37 - 33,47)^2}{33,47} + \frac{(46 - 50,20)^2}{50,20} + \ldots + \frac{(9 - 9,85)^2}{9,85} \approx 0,85$$

Puisque le système complet d'événements utilisé comporte 5 événements et qu'il a fallu procéder à l'estimation d'un paramètre avant de calculer les effectifs théoriques, le nombre de degrés de liberté de la loi du χ^2 gouvernant le comportement de cette distance sous (H_0) est $\nu = 5 - 1 - 1 = 3$.
Ensuite, pour une variable X suivant une loi du χ^2 à $\nu = 3$ degrés de liberté, on a $\mathbb{P}\{|X| > q_{0,05}\} = 0,05$ pour $q_{0,05} = 7,815$.
Comme $0,85 < 7,815$, on est largement en mesure d'accepter l'hypothèse (H_0) au risque $\alpha = 5\%$.

6.3.2 Tests d'homogénéité : comparaison de plusieurs distributions

I- Situation expérimentale et hypothèse de référence :
On considère une population de n individus partagée en l sous-populations : la première sous-population comporte N_1 individus, la seconde N_2 individus, ... , la dernière N_l individus. Notre but est de savoir si l'on peut considérer que ces différents groupes (ou sous-populations) exhibent un seul et même type de comportement relativement à une certaine variable. On considère donc à nouveau un système complet d'événements $A_1, A_2, \ldots, A_k$ (correspondant en général aux valeurs possibles d'une certaine variable), mais cette fois-ci les effectifs associés à chaque événements sont relevés à l'intérieur de chacune des l sous-populations : l'événement A_i est réalisé par $n_{i,1}^{obs.} = Eff._{i,1}^{obs.}$ individus du 1er groupe (1ère sous-population), mais aussi par $n_{i,2}^{obs.} = Eff._{i,2}^{obs.}$ individus du 2ème groupe, ... , et par $n_{i,l}^{obs.} = Eff._{i,l}^{obs.}$ individus du $l$ème groupe.
L'hypothèse (H_0) affirme que pour chaque indice $i \in [\![1; k]\!]$, les probabilités d'observer l'événement A_i sont identiques à l'intérieur de chacun des groupes considérés.

On dispose donc d'un premier tableau d'*"effectifs observés"* dans chacun des groupes et d'un second tableau d'*"effectifs théoriques"*. Evénement par événement, les effectifs théoriques sont proportionnels à la taille du groupe considéré : on a

$$Eff._{i,1}^{th.} = N_1 \cdot \frac{1}{n} \sum_{j=1}^{l} Eff._{i,j}^{obs.},\ Eff._{i,2}^{th.} = N_2 \cdot \frac{1}{n} \sum_{j=1}^{l} Eff._{i,j}^{obs.}, \ldots, Eff._{i,l}^{th.} = N_l \cdot \frac{1}{n} \sum_{j=1}^{l} Eff._{i,j}^{obs.},$$

Il s'agit à nouveau de comparer les tableaux *observé* et *théorique* en évaluant une certaine distance les séparant, puis d'accepter l'hypothèse (H_0) si ces deux tableaux sont "raisonnablement proches" (faible distance) ou de rejeter l'hypothèse (H_0) si ces deux tableaux sont "plutôt éloignés" (distance élevée).

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}}$$

suit approximativement une loi du χ^2 à $\nu = (k-1) \times (l-1)$ degrés de liberté.

Attention cependant : une fois encore, on exigera des événements $A_1, A_2, \ldots, A_k$ qu'ils soient suffisamment importants pour que les effectifs théoriques $Eff._{i,j}^{th.}$ soient tous ≥ 5. Si tel n'est pas le cas, il faut reprendre ce calcul de distance en regroupant les événements de moindre importance, dans le but d'obtenir un système complet dans lequel tous les effectifs théoriques sont ≥ 5.

III- Valeur à calculer :
il nous faut donc calculer la *"Distance du χ^2"*

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}}$$

puis rejeter l'hypothèse de référence (H_0) si la distance obtenue est "anormalement élevée". Ici encore, bien entendu, la distance du χ^2 à évaluer pourrait être calculée à partir des *fréquences observées* et des *fréquences théoriques* sous l'hypothèse (H_0) : puisqu'une fréquence s'obtient à partir d'un effectif en le divisant par n, on a

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = n\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{i,j}^{obs.} - f_{i,j}^{th.})^2}{f_{i,j}^{th.}}$$

IV- Définition de la Zone de Rejet :
une fois cette distance calculée, il reste à aller rechercher la valeur du quantile q_α associé à une variable X suivant une loi du χ^2 à $\nu = (k-1)\times(l-1)$ degrés de liberté : le quantile q_α étant tel que $\mathbb{P}\{X > q_\alpha\} = \alpha$, on rejette l'hypothèse nulle de conformité si la distance obtenue est supérieure à q_α, et on accepte cette hypothèse (H_0) sinon.

Voici un exemple d'application de ces tests d'homogénéité :

Exemple 10 : comparaison des effets de deux traitements
Les résultats de l'évolution d'une maladie suite à l'emploi de l'un ou l'autre des traitements A ou B figurent dans le tableau ci-dessous, ainsi que le nombre total de malades ayant suivi le traitement A ou le traitement B :

	Guérison	***Amélioration***	***Stationnaire***	***Totaux***
Traitement A	***280***	***210***	***110***	***600***
Traitement B	***220***	***90***	***90***	***400***
Totaux	***500***	***300***	***200***	***1000***

Au risque 5%, peut-on affirmer que les deux traitements produisent le même effet ?

Augmentons le tableau précédent en calculant, événement par événement et traitement par traitement, des effectifs théoriques sous l'hypothèse (H_0) où les deux traitements produisent des effets identiques :

	Guérison	*Amélioration*	*Stationnaire*	*Totaux*
Traitement A	***280 / 300***	***210 / 180***	***110 / 120***	***600***
Traitement B	***220 / 200***	***90 / 120***	***90 / 80***	***400***
Totaux	***500***	***300***	***200***	***1000***

Puisque tous les effectifs théoriques obtenus sont supérieurs à cinq, on peut calculer directement la distance du χ^2 séparant le tableau empirique du tableau théorique, ce qui nous donne ici :

$$V = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}} = \frac{(280-300)^2}{300} + \frac{(210-180)^2}{180} + \ldots + \frac{(90-80)^2}{80} \approx 17,92$$

Ici, le nombre de degrés de liberté de la loi du χ^2 gouvernant le comportement de cette distance sous (H_0) est $\nu = (k-1)(l-1) = (3-1)(2-1) = 2$.
Ensuite, pour une variable X suivant une loi du χ^2 à $\nu = 2$ degrés de liberté, on a $\mathbb{P}\{|X| > q_{0,05}\} = 0,05$ pour $q_{0,05} = 5,991$.
Comme $17,92 > 5,991$, on peut rejeter l'hypothèse (H_0) au risque $\alpha = 5\%$.
Cette hypothèse d'homogénéité peut même être rejetée au risque très faible de $\alpha' = 0,1\%$, puisque pour une variable suivant une loi du χ^2 à deux degrés de liberté : $q_{0,001} = 13,815$.
Il convient donc de considérer que ces deux traitements produisent des effets différents.

6.3.3 Tests d'indépendance entre deux variables

I- Situation expérimentale et hypothèse de référence :
Considérons enfin un premier système complet d'événements $A_1, A_2, \ldots, A_k$ correspondant aux valeurs possibles d'une certaine variable U, mais aussi un deuxième système complet d'événements $B_1, B_2, \ldots, B_l$ correspondant aux valeurs possibles d'une autre variable W.
Dans une population de n individus, on relève combien de fois les événements A_i et B_j ont tous deux été réalisés, ce qui nous donne un tableau d'effectifs observés $n_{i,j}^{obs.}$ où $1 \leq i \leq k$ et $1 \leq j \leq l$.
L'hypothèse (H_0) affirme que les variables U et W sont indépendantes.

On pourra donc constituer un second tableau d'*"effectifs théoriques"* correspondant à des fréquences théoriques elles-mêmes obtenues par une *règle du produit* :

$$f_{i,j}^{th.} = f_{i,.}^{obs.} \times f_{.,j}^{obs.} \text{ et donc } n_{i,j}^{th.} = n \times f_{i,.}^{obs.} \times f_{.,j}^{obs.},$$

les fréquences marginales $f_{i,.}$ et $f_{.,j}$ étant naturellement données par $f_{i,.}^{obs.} = \sum_j f_{i,j}^{obs.}$ et $f_{.,j}^{obs.} = \sum_i f_{i,j}^{obs.}$.

Il s'agit encore de comparer les tableaux *observé* et *théorique* en évaluant une certaine distance les séparant, puis d'accepter l'hypothèse (H_0) si ces deux tableaux sont "raisonnablement proches" (faible distance) ou de rejeter l'hypothèse (H_0) si ces deux tableaux sont "plutôt éloignés" (distance élevée).

II- Résultat théorique à utiliser :
sous (H_0), la variable aléatoire

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}}$$

suit approximativement une loi du χ^2 à $\nu = (k-1)\times(l-1)$ degrés de liberté.

Attention cependant : une fois encore, on exigera des événements $A_1, A_2, \ldots, A_k$ et $B_1, B_2, \ldots, B_l$ qu'ils soient suffisamment importants pour que les effectifs théoriques $Eff._{i,j}^{th.}$ soient tous ≥ 5. Si tel n'est pas le cas, il faut reprendre ce calcul de distance en regroupant les événements de moindre importance, dans le but d'obtenir un système complet dans lequel tous les effectifs théoriques sont ≥ 5.

III- Valeur à calculer :
il nous faut donc calculer la *"Distance du χ^2"*

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}}$$

puis rejeter l'hypothèse de référence (H_0) si la distance obtenue est "anormalement élevée". Ici encore, bien entendu, la distance du χ^2 à évaluer pourrait être calculée à partir des *fréquences observées* et des *fréquences théoriques* sous l'hypothèse (H_0) : puisqu'une fréquence s'obtient à partir d'un effectif en le divisant par n, on a

$$V = D_{\chi^2}(Tab.Obs.; Tab.Th.) = n\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(f_{i,j}^{obs.} - f_{i,j}^{th.})^2}{f_{i,j}^{th.}}$$

IV- Définition de la Zone de Rejet :
une fois cette distance calculée, il reste à aller rechercher la valeur du quantile q_α associé à une variable X suivant une loi du χ^2 à $\nu = (k-1)\times(l-1)$ degrés de liberté : le quantile q_α étant tel que $\mathbb{P}\{X > q_\alpha\} = \alpha$, on rejette l'hypothèse nulle d'indépendance si la distance obtenue est supérieure à q_α, et on accepte cette hypothèse (H_0) sinon.

Voici donc un exemple d'application de ces tests d'indépendance :

Exemple 11 : âges des patients et effet d'un traitement
A la suite d'un même traitement, on a observé 40 bons résultats chez 70 malades jeunes et 50 bons résultats chez une centaine de malades âgés. Au risque 10%, peut-on affirmer qu'il existe une liaison entre l'âge du malade et l'effet du traitement ?

Augmentons le tableau précédent en y introduisant des effectifs théoriques sous l'hypothèse (H_0) où l'effet du traitement est indépendant de l'âge du patient :

	Bons résultats	***Mauvais résultats***	***Totaux***
Patients jeunes	***40 / 37,06***	***30 / 32,94***	***70***
Patients agés	***50 / 52,94***	***50 / 47,06***	***100***
Totaux	***90***	***80***	***170***

Puisque tous les effectifs théoriques obtenus sont supérieurs à cinq, on peut calculer directement la distance du χ^2 séparant le tableau empirique du tableau théorique, ce qui donne :

$$V = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(Eff._{i,j}^{obs.} - Eff._{i,j}^{th.})^2}{Eff._{i,j}^{th.}} = \frac{(40-37,06)^2}{37,06} + \frac{(30-32,94)^2}{32,94} + \frac{(50-52,94)^2}{52,94} + \frac{(50-47,06)^2}{47,06} \approx 0,84$$

Ici, le nombre de degrés de liberté de la loi du χ^2 gouvernant le comportement de cette distance sous (H_0) est $\nu = (k-1)(l-1) = (2-1)(2-1) = 1$.
Ensuite, pour une variable X suivant une loi du χ^2 à un seul degré de liberté, on a $\mathbb{P}\{|X| > q_{0,1}\} = 0,1$ pour $q_{0,1} = 2,706$.
Comme $0,84 < 2,706$, on ne peut pas rejeter l'hypothèse (H_0) d'indépendance au risque $\alpha = 10\%$.

6.4 Exercices d'Application

I Echantillonnage et Tests d'Efficacité :

1. Au niveau national, en 2014, les candidats aux trois types de baccalauréat se sont répartis de la façon suivante: général 47% ; technologique 20% ; professionnel 33%.
   Sur un échantillon de 250 candidats, on a observé la répartition d'effectifs suivante : général 131 ; technologique 37 ; professionnel 82.
   Cet échantillon est-il représentatif de la population nationale? (*Vous répondrez en effectuant un test au niveau de risque 5% puis 10%.*)
2. Les tableaux suivants proviennent d'une enquête effectuée auprès de jeunes ayant subi les épreuves du baccalauréat général ou technologique.

Bac général

	Présents	Admis
Garçons	150	132
Filles	186	173

Bac technologique

	Présents	Admis
Garçons	70	62
Filles	72	67

   (a) Transformer le tableau concernant le bac général pour qu'il se présente sous la forme d'un tableau croisé des variables "Sexe" et "Réussite", prenant les valeurs respectives $\{G, F\}$ et $\{Admis, Pas\ admis\}$.

   (b) Effectuer un test au niveau de risque 5% pour décider si, au vu de l'échantillon, la phrase suivante est fondée: "Au bac général, les filles réussissent mieux que les garçons."
   (*Indic: on trouve une distance du khi-deux égale 2.49*)

   (c) En ce qui concerne le bac technologique, le même test conduit à une valeur de khi-deux égale à 0.86 (on ne demande pas de faire ce calcul). Pour quel type de bac, général ou technologique, la réussite est-elle plus conditionnée au sexe? (*justifiez votre réponse*)

   (d) Les chiffres des tableaux correspondent **en milliers** aux résultats nationaux du bac 2014. Que répondez-vous la question posée en **b)** en tenant compte de toute la population?
   e)* Utiliser les deux tableaux de l'enquête pour tester l'indépendance entre le sexe et le choix de la filière technologique ou générale.

II Dé électronique ⊞ :

On considère un dé électronique à 6 faces, numérotées de 1 à 6. On a "lancé" le dé 1000 fois et obtenu les résultats suivants

face	1	2	3	4	5	6
effectif	155	170	155	152	168	200

1. A votre avis, ce dé est-il équilibré? *Vous répondrez en effectuant un test d'ajustement au niveau de risque de votre choix.*
2. Auriez-vous fourni pareille réponse si les mêmes proportions avaient été observées sur 10000 lancers ?
3. Tester l'hypothèse selon laquelle le dé est pipé en faveur du 6 avec probabilité $\frac{1}{6} + \frac{1}{10}$, les autres faces étant équiprobables entre elles.
4. ⊞ Simuler 1000 lancers de 2 dés à 6 faces équilibrés. Afficher les résultats obtenus pour la somme des deux dés dans un tableau d'effectifs. Effectuer un test d'ajustement avec la loi théorique attendue.

III Tests de générateurs aléatoires ⊞ :

Pour générer des nombres aléatoires, qu'ils soient entiers ou réels, `Scilab` dispose de fonctions prédéfinies qui produisent des lois précises. Voici quelques exemples que vous trouverez, ainsi que d'autres, dans la rubrique `grand` (prononcez "g-rand") de l'aide.
- lois discrètes pour générer des nombres entiers : loi Binomiale, loi Uniforme dans $\{1, 2, \ldots, n\}$, loi de Poisson,
- lois continues pour générer des nombres réels : loi Uniforme dans $[a, b]$, loi Normale, loi Exponentielle.

1. Choisir une loi discrète, et les paramètres nécessaires, parmi celles proposées ci-dessus et simuler un échantillon de taille 1000 de cette loi.
   Calculer la moyenne de l'échantillon et comparer avec la moyenne attendue.
   Tracer un histogramme de la répartition de l'échantillon.
2. Construire une fonction `testkhi2` qui permet d'effectuer un test d'ajustement. Cette fonction prendra comme variables d'entrée `(obs,theo,niv)` et fournira en variables de sortie `[dist,concl]`, où
   - `obs,theo` désignent deux vecteurs lignes de même longueur contenant les effectifs observés et les effectifs théoriques dans un tableau d'effectifs,
   - `niv` désigne le niveau du test à fixer entre 0 et 1,
   - `dist` désigne la distance du khi-deux entre le tableau observé et le tableau théorique,
   - `concl` désigne la conclusion du test d'ajustement sous la forme `rejet` ou `pas rejet`.
3. Utiliser la fonction de la question b) pour effectuer un test d'ajustement sur les données simulées à la question a).
4. Reprendre les questions a) et c) avec une loi uniforme dans $[0, 1]$, puis avec une loi uniforme dans $[a, b]$ (à vous de choisir a et b), puis (*) avec une autre loi continue.

IV Virus :

On désire tester la polyvalence d'un traitement sur une maladie dont le virus présente plusieurs souches à la suite de mutations génétiques. On effectue donc des tests sur des patients volontaires atteints par différentes souches, ce qui donne les résultats suivants :

Souche	1	2	3	4	5	6
Nb. malades	575	988	2240	209	210	287
Nb. guérisons	48	89	180	17	18	13

Tester au niveau $\alpha = 5\%$ l'hypothèse : "la souche n'intervient pas dans les chances de guérison". *Attention : effectuer un test d'indépendance entre deux variables clairement identifiées et adapter le tableau en conséquence!*

V Echantillonnage et Tests d'Efficacité :

On sait qu'une maladie atteint 10% des jeunes ovins d'une région donnée. Un chercheur a expérimenté un traitement sur un échantillon de n agneaux. Il a recensé 5% d'agneaux malades sur cet échantillon.
Quelle est la valeur minimale de l'entier n (taille de l'échantillon d'agneaux traités) permettant de conclure, au risque $\alpha = 5\%$, à une certaine efficacité du traitement appliqué ?

VI To smoke ... or not to smoke :

Après avoir suivi pendant vingt ans un groupe de 200 sujets, on a observé les résultats suivants en comptabilisant les cancers apparus dans chacun des deux groupes :

	Non–fumeurs	***Fumeurs***
Apparition d'un cancer	***20***	***40***
Pas de cancer	***180***	***160***

Utiliser deux méthodes de test différentes pour décider, au risque $\alpha = 5\%$, si les différences observées entre fumeurs et non-fumeurs sont significatives ou non.

VII Réussite au Baccalauréat :

Le taux de réussite au Baccalauréat d'une certaine série, une année donnée, est de 67%.
Pour les deux questions qui suivent, on effectuera des tests au niveau de risque $\alpha = 5\%$.

1. Dans le centre d'examen A, il y a eu 216 reçus pour 300 candidats présents. Les résultats de ce centre sont-ils conformes aux résultats nationaux ?
2. Dans un centre d'examen B situé dans la même ville, il y a eu 128 reçus pour 200 candidats. Les résultats des centres A et B sont-ils significativement différents ?

VIII Botanique Mendélienne :

On a effectué le croisement de balsamines blanches avec des balsamines pourpres. En première génération, les fleurs sont toutes pourpres. On obtient en deuxième génération quatre catégories, avec les effectifs suivants :

Couleurs	*Pourpre*	*Rose*	*Blanc lavande*	*Blanc*
Effectifs	*1790*	*547*	*548*	*213*

Au risque $\alpha = 5\%$, peut-on accepter l'hypothèse de répartition mendélienne $(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$?

IX Groupes sanguins et apparitions d'une maladie :
On cherche à savoir si la fréquence d'apparition d'une maladie est liée au groupe sanguin. Sur 200 malades observés, on a dénombré 104 personnes du groupe O, 76 du groupe A, 18 du groupe B et 2 du groupe AB.
On admet que dans la population générale, la répartition entre les groupes est la suivante : 47% d'individus dans le groupe O, 43% dans le groupe A, 7% dans le groupe B et 3% dans le groupe AB. Que peut-on en conclure ?

X Hypothèse d'une répartition gaussienne :
Lors d'une étude biologique portant sur une certaine espèce de mollusque, on a mesuré le taux de protéines x en mg de 36 individus de cette espèce, obtenant les résultats suivants :

x	*]0 ; 1,5]*	*]1,5 ; 3]*	*]3 ; 4,5]*	*]4,5 ; 6]*	*]6 ; 7,5]*	*]7,5 ; 9]*	*]9;10,5]*
Nombre d'individus	*8*	*7*	*4*	*9*	*2*	*3*	*3*

1. Donner des estimations non-biaisées de la moyenne et de l'écart-type associés à la variable x au sein de cette population.
2. Peut-on admettre l'hypothèse selon laquelle ce taux est distribué de façon gaussienne au sein de la population considérée ?

XI Test Poissonien dans une situation réelle Des plaignants[6] ont poursuivi en justice le Ministère israélien de la Santé suite à une campagne de vaccination menée sur des enfants et ayant entraîné des dommages fonctionnels irréversibles pour certains d'entre eux. Ce vaccin était en fait connu pour entraîner de tels dommages en de très rares circonstances ;

[6] cf. *Murray Atkin*, "Evidence and the Posterior Bayes Factor", 17 Math. Scientist 15 (1992)

des études antérieures menées dans d'autres pays ont montré que ce risque était d'un cas sur 310'000 vaccinations. Les plaignants avaient été informés de ce risque et l'avaient accepté. Les doses ayant provoqué les dommages objet de la plainte provenaient d'un lot ayant servi à vacciner un groupe de 300'533 enfants ; dans ce groupe, quatre cas de dommages ont été diagnostiqués.

1. On modélise l'événement *"Le vaccin provoque des dommages fonctionnels irréversibles sur l'enfant i"* par une variable aléatoire de Bernoulli X_i de paramètre p. Calculer la valeur p_0 correspondant aux résultats des études antérieures.
2. Montrer que l'on peut modéliser la loi du nombre N de cas de dommages dans le groupe par une loi de Poisson de paramètre λ. Calculer la valeur λ_0 attendue sous l'hypothèse d'une conformité du vaccin aux études antérieures.
3. L'hypothèse (H_0) : "$p = p_0$" correspond au risque que les plaignants avaient accepté d'encourir, l'hypothèse alternative étant (H_1) : "$p > p_0$".
   Construire, à partir de la variable N, un test de niveau α pour l'hypothèse (H_0). Accepte-t-on (H_0) pour un niveau de risque $\alpha = 5\%$?

XII Locations et saisons :

Des appartements à la montagne peuvent être loués à la semaine. Dans la comparaison des taux d'occupation de ces appartements pour un mois d'hiver (Janvier) et pour un mois d'été (Juillet), on dispose de deux échantillons, l'un de 300 observations instantanées en Janvier, l'autre de 200 observations instantanées en Juillet :

	Janvier	*Juillet*
Occupation	***240***	***150***
Inoccupation	***60***	***50***

Au risque $\alpha = 5\%$, convient-il de considérer que les taux d'occupation de ces appartements sont identiques en Janvier et en Juillet ?

XIII Estimation de Taux de Fertilité :

On considère un échantillon de 169 brebis de la race "Ile de France". Ces brebis ont été mises en lutte, on a obtenu 108 brebis pleines (c'est à dire fécondées).
A l'issue de cette expérience, on cherche à estimer le taux de fertilité t de la race "Ile de France".
QCM1 : Soit $\hat{t}$ l'estimateur de t obtenu par la méthode du maximum de vraisemblance. Peut-on affirmer que

A $\hat{t}$ est un estimateur sans biais de t ?

B $\hat{t}$ est un estimateur biaisé de t ?

C $\hat{t} = 69,2\%$?

D $\hat{t} = 63,9\%$?

E $\hat{t} = 59,2\%$?

QCM2 : Un intervalle de confiance au niveau 95% pour ce taux t est donné par

A $I = [0,636; 0,651]$?

B $I = [0,586; 0,651]$?

C $I = [0,566; 0,721]$?

D $I = [0,546; 0,731]$?

E $I = [0,586; 0,601]$?

QCM3 : Peut-on affirmer que

A Un intervalle de confiance au niveau 99% pour t sera plus large que celui obtenu au niveau 95% ?

B Un intervalle de confiance au niveau 90% pour t sera plus large que celui obtenu au niveau 95% ?

C Un intervalle de confiance au niveau 99% pour t sera plus étroit que celui obtenu au niveau 95% ?

D Avec 200 brebis plutôt que 169, même si l'estimée ponctuelle de t devait différer de celle obtenue précédemment, on serait en mesure de fournir un intervalle de confiance plus étroit pour t au niveau 95% ?

E Avec 500 brebis plutôt que 169, si l'estimée ponctuelle de t est proche de celle obtenue précédemment, on sera en mesure de fournir un intervalle de confiance plus étroit pour t au niveau 95% ?

XIV Taux d'échec d'un vaccin :

Des études antérieures ont révélé que le taux d'échec τ associé à une certaine vaccination est situé entre 10% et 15%.

On prépare une expérience dans le but de déterminer à 1% près la proportion de sujets non-immunisés par ce vaccin, en acceptant un coefficient de risque $\alpha = 0,05$. n personnes seront donc vaccinées, et l'on relèvera par la suite qui parmi ces n personnes n'est pas immunisé à l'issue de la vaccination.

QCM4 : Peut-on affirmer que

A Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit une loi de Poisson ?

B Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit une loi binomiale ?

C Le nombre de sujets non-immunisés à l'issue de cette nouvelle campagne suit approximativement une loi de Poisson ?

D Les variables "nombre de sujets non-immunisés" et "nombre de sujets immunisés" sont indépendantes ?

E Les variables "nombre de sujets non-immunisés" et "nombre de sujets immunisés" sont négativement corrélées ?

QCM5 : Pour parvenir à ses fins (déterminer à 1% près et au niveau de confiance 95% la proportion de sujets non-immunisés), quel est le nombre minimal n_0 de sujets que l'on doit vacciner ?

A $n_0 = 100$

B $n_0 = 1000$

C $n_0 = 3900$

D $n_0 = 4900$

E $n_0 = 5400$

XV Encore des sondages :

A la veille d'une consultation électorale majeure, on a interrogé une centaine d'électeurs constituant un échantillon représentatif. 58 d'entre eux ont déclaré avoir l'intention de voter pour le candidat *Toulemonde*.

QCM6 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A $I = [0,453; 0,707]$ constitue un intervalle de confiance au niveau 95% pour la cote de M. Toulemonde.

B $I = [0,483; 0,677]$ constitue un intervalle de confiance au niveau 95% pour la cote de M. Toulemonde.

C $I = [0,453; 0,707]$ constitue un intervalle de confiance au niveau 99% pour la cote de M. Toulemonde.

D Au niveau de confiance 99%, on peut affirmer que M. Toulemonde va remporter ces élections.

E Au niveau de confiance 99%, on n'est pas en mesure de garantir que M. Toulemonde va remporter ces élections.

QCM7 : Pour une même fréquence observée d'électeurs favorables à M. Toulemonde, quelle devrait être la taille minimale n_0 de l'échantillon permettant d'affirmer au niveau de confiance 95% que M. Toulemonde sera élu ?

A $n_0 = 100$

B $n_0 = 103$

C $n_0 = 147$

D $n_0 = 200$

E $n_0 = 10000$

XVI Estimations de valeurs moyennes au sein de deux populations :

Dans une région d'Europe, l'étude de la masse du cerveau mesurée en grammes chez des sujets âgés de 20 à 49 ans a conduit aux résultats suivants :

Hommes

Val. approx.	***1170***	***1220***	***1270***	***1320***	***1370***	***1420***	***1470***	**Total**
Effectifs	***5***	***36***	***45***	***50***	***61***	***49***	***19***	***265***

Femmes

Val. approx.	***1070***	***1120***	***1170***	***1220***	***1270***	***1320***	***1370***	**Total**
Effectifs	***12***	***22***	***45***	***54***	***52***	***20***	***10***	***215***

QCM8 : Quel est l'intervalle de confiance au niveau 99% pour la valeur moyenne de cette masse au sein de la population masculine ?

A $I = [1326; 1345]$

B $I = [1323; 1349]$

C $I = [1328; 1344]$

D $I = [1322; 1350]$

E $I = [1320; 1351]$

QCM9 : Quel est l'intervalle de confiance au niveau 99% pour la valeur moyenne de cette masse au sein de la population féminine ?

A $I' = [1203; 1230]$

B $I' = [1209; 1233]$

C $I' = [1209; 1229]$

D $I' = [1206; 1233]$

E $I' = [1206; 1330]$

QCM10 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A Pour l'ensemble de la population, une estimation ponctuelle de cette masse moyenne s'obtient en effectuant la demi-somme $\frac{1}{2}(\hat{m}_h + \hat{m}_f)$.

B Pour l'ensemble de la population, une estimation ponctuelle de cette masse moyenne s'obtient en effectuant une somme $\frac{265}{480}\hat{m}_h + \frac{215}{480}\hat{m}_f$.

C Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est moins large que pour la seule population masculine ?

D Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est plus large que pour la seule population masculine ?

E Pour l'ensemble de la population, l'intervalle de confiance au niveau 99% pour cette valeur moyenne est de même largeur que pour la seule population masculine ?

XVII Estimation de la fréquentation d'un hôpital :

On considère un hôpital comportant cent salles de consultation. Chacune de ces salles accueille quotidiennement un nombre de patients qui est modélisé par une variable de Poisson de paramètre $\lambda = 10$, et l'on suppose que ces variables de Poisson sont indépendantes. Soit S le nombre total de patients qui viennent pour une consultation à l'hôpital un jour donné, puis $M = S/100$ le nombre moyen de consultations par salle ce jour-là.

QCM11 : Quelles sont les propositions correctes parmi celles ci-dessous ?

A $\mathbb{E}(M) = 0,1$

B $\mathbb{E}(M) = 10$

C $\mathbb{E}(S) = 100$

D $\mathbb{E}(S) = 1000$

E $\mathrm{Var}(S) = 10000$

QCM12 : Quelle est la probabilité de voir S dépasser la valeur 1050 un jour donné (à 10^{-2} près) ?

A $0,06$

B $0,11$

C $0,26$

D $0,48$

E $0,96$

XVIII Estimation d'un paramètre poissonien :

On suppose que le nombre d'accidents survenant dans une certaine ville un jour donné suit une loi de Poisson de paramètre λ inconnu. Pendant un an, on a relevé quotidiennement le nombre X d'accidents survenus dans cette ville durant la journée, le tableau ci-dessous résume les résultats de cette enquête :

Nombre d'accidents dans la journée

Val. observ.	***0***	***1***	***2***	***3***	***Total***
Nombre de j.	***200***	***100***	***55***	***10***	***365***

On part du principe que les variables $X_1, X_2, \ldots, X_{365}$ associées à chaque journée sont indépendantes, et l'on note M la variable aléatoire de moyenne annuelle des nombres d'accidents quotidiens.

QCM13 : Peut-on affirmer que

A M suit exactement une loi de Poisson ?

B M suit approximativement une loi binomiale ?

C M suit approximativement une loi normale ?

D $\mathbb{E}(M) = \lambda$?

E $\mathrm{Var}(M) = \frac{\lambda}{365}$?

QCM14 : On note s^2 la variance empirique obtenue pour ces variables de Poisson $X_1, X_2, \ldots, X_{365}$ et $\hat{x}$ leur moyenne empirique. Peut-on affirmer que

A $\hat{x} = 0,66$?

B $\hat{x} = 91,3$?

C $s^2 = 0,69$?

D $s^2 = 6,85$?

E $s^2 = 68,5$?

QCM15 : On choisit de noter $I_{1-\alpha}$ un intervalle de confiance de niveau $(1-\alpha)$ pour le paramètre λ. En arrondissant les bornes à 10^{-2} près, peut-on affirmer que

A $I_{0,95} = [0,57; 0,75]$?

B $I_{0,95} = [0,64; 0,68]$?

C $I_{0,95} = [90,5; 92,1]$?

D $I_{0,90} = [0,59; 0,73]$?

E $I_{0,90} = [90,8; 91,8]$?

XIX Mesures et précision :

Un appareil dose la concentration d'une substance sans biais, mais avec une erreur de mesure suivant une loi normale de moyenne nulle et de variance $100\,(mg/L)^2$. Si on effectue plusieurs mesures d'une concentration sur un même prélèvement, on suppose que toutes les erreurs de mesures sont indépendantes. En répétant les mesures puis en choisissant leur moyenne comme résultat, on souhaite obtenir une précision (demi-longueur d'int. de confiance de niveau 95%) inférieure ou égale à $5\,mg/L$.

QCM16 : Combien faut-il effectuer de mesures pour obtenir la précision désirée ?

A 5

B 10

C 12

D 16

E 20

QCM17 : Avec l'appareil de la question précédente, le coût d'une mesure est de 100 Euros. On se demande s'il est rentable d'acheter un nouvel appareil pour doser la concentration de la substance. Un nouvel appareil produit des mesures sans biais et comportant une erreur de variance $40\,(mg/L)^2$ seulement, mais le coût de chaque mesure avec ce nouvel appareil s'élève à 200 Euros. Ici encore, l'objectif est d'obtenir une précision inférieure ou égale à $5\,mg/L$, quitte à combiner de nombreuses mesures. Peut-on affirmer que

A Pour atteindre ce but, il faut effectuer 5 mesures avec le nouvel appareil ?

B Pour atteindre ce but, il faut effectuer 7 mesures avec le nouvel appareil ?

C Pour atteindre ce but, il faut effectuer 9 mesures avec le nouvel appareil ?

D Il vaut mieux garder l'ancien appareil de dosage.

E Il vaut mieux acheter le nouvel appareil de dosage.

XX Encore des sondages :

Une élection oppose deux candidats, et tout porte à croire qu'elle sera très serrée : on part du principe que tous les électeurs vont voter pour l'un ou l'autre candidat, et que les scores de chacun d'eux seront proches de 50%. On souhaite effectuer un sondage à la sortie des bureaux de vote.

QCM18 : Combien d'électeurs faut-il interroger pour atteindre une précision (demi-longueur d'intervalle de confiance au niveau 95%) de $0,02$?

A 996

B 1010

C 1563

D 2401

E 2542

QCM19 : Si l'on choisit d'interroger 1000 électeurs, quelle sera la précision obtenue pour les proportions de voix remportées par chaque candidat ?

A $0,01$

B $0,03$

C $0,05$

D $0,07$

E $0,10$