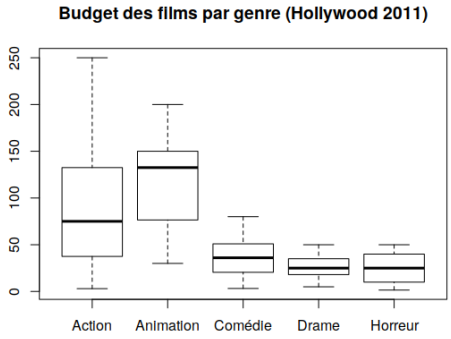
Statistique descriptive

Exercice 1. Soit le graphique suivant :



1. À quel genres appartiennent les films de petit budget ? Quel genre de film a le budget le moins variable ?
2. 25% des films d'animation coûte plus que ...
3. Quelle proportion de films d'action a un budget supérieur au budget médian des comédies ?

♣ *Réponse: Pour la première question : c'est plutôt «drame comédie, horreur» ; drame et horreur ne se distinguent que par leur variabilité. C'est «drame» qui a la moindre variabilité.*

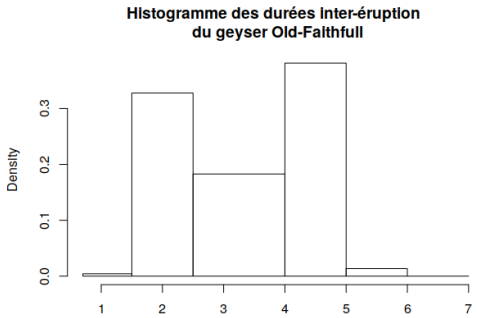
Exercice 2. Pour étudier le nombre d'enfants de moins de 18 ans par famille, on choisit un échantillon de familles et pour chacune d'elles, on note le nombre d'enfants. La répartition des familles de l'échantillon suivant le nombre d'enfants est donnée par le tableau :

nombre d'enfants	0	1	2	3	4	5	6	7	8
nombre de familles	91	146	104	63	47	33	10	4	2

1. Construire le diagramme bâtons de ces données.
2. Calculer le nombre moyen d'enfants par famille dans l'échantillon.
3. Construire le tableau qui à chaque n associe le nombre de familles ayant au plus n enfants de moins de 18 ans. Donner la médiane et les quartiles du nombre d'enfants de moins de 18 ans.

♣ *Réponse: Moyenne*$= 2$. *Tableau :* 91 237 341 404 451 484 494 498 500. *Les quantiles sont associées au* $125^{\text{è}}$, $250^{\text{è}}$, $375^{\text{è}}$ *échantillon, soit* $Q_1 = 1, Q_2 = 2, Q_3 = 3$.

Exercice 3. Soit l'histogramme suivant basé sur 300 individus :



Combien de données ont été mesurées entre 2 minutes 30 et 4 minutes.

♣ *Réponse:* $1.5 \times 0.2 = n/300 \ \rightarrow \ n = 90$.

Exercice 4. Les salaires annuels des 30 employés d'une entreprise sont les suivants (en centaines d'euros), présentés par ordre croissant :

100	110	110	120	130	130	150	150	150	150	180	180	180	180	180
180	220	220	220	220	220	220	250	250	260	340	360	440	440	560

1. Donner la médianes et les quartiles de cette série.
2. Calculer la moyenne.
3. Tracer l'histogramme en regroupant les données en classes successives de longueur 100, 100, 200, et 100 la première étant l'intervalle [100,200[. Rappel : La surface est la proportion.

♣ *Réponse: $Q_2 = 180$, $Q_1 = 150$, $Q_3 = 250$. Le calcul de la somme est un exercice de calcul mental : on peut regrouper en sous paquets dont la somme est un multiple de 100. La somme fait 6600. Pour l'histogramme, la solution complète est :*

intervalle	$[100, 200[$	$[200, 300[$	$[300, 500[$	$[500, 600[$
n_i	*16*	*9*	*4*	*1*
l_i	*100*	*100*	*200*	*100*
$30 \times h_i$	*0,16*	*0,09*	*0,02*	*0,01*

.

Exercice 5. Soit $(x_1, ...x_n)$ une suite de données numériques. Notons $\bar{x}$ et s_x les moyennes et écarts type associés.

1. Soit a, b deux réels, que valent les moyennes et écarts type des suites $y_i = x_i - a$ et $z_i = x_i/b$?
2. Soit a, b deux réels, que valent les moyennes et écarts type de la suites $t_i = (x_i - a)/b$?
3. Que valent la moyenne et l'écart type des suites $(x_i - \bar{x})$ et $(x_i - \bar{x})/s_x$?
4. Soit a réel. Exprimer $\frac{1}{n}\sum x_i^2$ en fonction de s_x et $\bar{x}$. En déduire une expression de $\frac{1}{n}\sum(x_i - a)^2$ en fonction de s_x et $a - \bar{x}$.
5. Pour convertir en degrés Celsius une température donnée en degrés Fahrenheit, il suffit de soustraire 32 et de multiplier par 5/9. La température moyenne à Rennes au mois de janvier 2019 à été de 5,5 degrés Celsius avec une variance de 16. Donner la moyenne et l'écart-type de ces températures en degrés Fahrenheit. On demande les valeurs numériques exactes.

♣ *Réponse: Le but est de faire manipuler des sommes, comprendre la linéarité, comprendre ce que signifie «centré-réduit». Dernière question : $m = 32 + \frac{9}{5} \times 5.5 = 41.9$, $\sigma = \frac{9}{5} \times 4 = 7,2$*

Exercice 6. Soit x un ensemble de données séparé en deux sous-ensembles y et z de taille n_y et n_z

1. Montrer que

$$\bar{x} = p_y\bar{y} + p_z\bar{z}, \quad p_y = \frac{n_y}{n_y + n_z}, \quad p_z = \frac{n_z}{n_y + n_z}$$

$$s_x^2 = \{p_y s_y^2 + p_z s_z^2\} + \{p_y(\bar{y} - \bar{x})^2 + p_z(\bar{z} - \bar{x})^2\}.$$

s_x^2 est donc la somme de deux termes, le premier étant la moyenne pondérée des variances, appelée variance intra-classe ; le second est appelé variance inter-classe.
Indication : Pour la seconde identité, on commencera par montrer que

$$\sum(y_i - \bar{x})^2 = \sum(y_i - \bar{y})^2 + n_y(\bar{y} - \bar{x})^2$$

en utilisant que $s_{y-\bar{x}} = s_y$.
♣ *Réponse: L'identité cherchée se réécrit $ns_{y-\bar{x}}^2 + n_y(\bar{y} - \bar{x})^2 = ns_y^2 + n_y(\bar{y} - \bar{x})^2$.*

2. Soit la variable aléatoire qui vaut $\bar{y}$ avec probabilité p_y et $\bar{z}$ avec probabilité p_z. Quelle est son espérance, sa variance ? Faire le lien avec ce qui précède.

Exercice 7. (Paradoxe de Simpson) On considère les statistiques suivantes sur les taux de réussites au baccalauréat de deux lycées :

	Lycée A	Lycée B	Total
Echecs	63	16	79
Réussites	2037	784	2821
Total	2100	800	2900
Taux d'échec	0,030	0,020	0,027

Quel lycée choisiriez-vous ? Une deuxième étude, plus fine, sépare les individus en deux groupes, ceux qui sont issus d'un milieu défavorisé et les autres :

	Favorisé			Défavorisé		
	Lycée A	Lycée B	Total	Lycée A	Lycée B	Total
Echecs	6	8	14	57	8	65
Réussites	594	592	1186	1443	192	1635
Total	600	600	1200	1500	200	1700
Taux d'échec	0,010	0,013	0,016	0,038	0,040	0,038

Quel lycée choisiriez-vous ? Expliquer brièvement le paradoxe en comparant la proportion d'enfant défavorisés dans les deux lycées.

♣ *Réponse: L'abondance de population défavorisée dans le lycée A fait qu'en dépit de ses meilleures performances dans chaque milieu, son taux d'échec global est proche du taux d'échec de la population défavorisée. On peut essayer de détailler davantage mais ça devient vite très embrouillé.*

Exercice 8. En 2007, le taux brut de mortalité en Inde est inférieur à celui de la France : 8 pour 1000 contre 9 pour 1000. Pourtant à tout âge le taux de mortalité est inférieur en France à ce qu'il est en Inde. Expliquer en s'inspirant de l'exercice précédent.

♣ *Réponse: Pareil : plein de jeunes en Inde et de vieux chez nous.*

Exercice 9. Soit la fonction $f(y) = |y| + |y-1| + |y-3|$. Calculer $f(0), f(1)$ et $f(3)$. Tracer cette fonction. On notera que cette fonction est continue, affine par morceaux, comme somme de trois fonctions de ce type.

Soit $(x_1, ...x_n)$ une suite de données numériques. Montrer que la médiane est la valeur m pour laquelle la somme des distances des données à cette valeur, i.e. $y \mapsto \sum_i |x_i - y|$, est minimale. On remarquera que la fonction $y \to \sum_i |x_i - y|$ est continue, affine par morceaux, avec une dérivée entière sur chaque morceau, qui augmente de 2 à chaque fois. On pourra commencer par traiter le cas où n est impair.

Statistiques à deux variables

Exercice 10. On considère les passagers du Titanic, avec les variables Survie, Sexe, et Classe :

	Femme			Homme		
	1re	2e	3e	1re	2e	3e
Survivant	134	94	80	59	25	58
Mort	9	13	132	120	148	441

1. Quelle est la probabilité de survie d'une femme de deuxième classe ? Même question pour un homme qui n'est pas en première classe.
2. Le prix du ticket était de 1000 $ en première, 500 $ en seconde, et 100 $ en troisième. Quel est le prix moyen du ticket d'un passager qui a survécu ?
3. Quel est le prix moyen du ticket d'un passager qui n'a pas survécu ?

Exercice 11. Voici des statistiques concernant le lien entre l'hypertension et la consommation d'alcool sur une population de 500 personnes[1]. La variable V représente le nombre de verres d'alcools consommés par jours, et H la présence d'hypertension.

H \ V	0	2	4	8
O(ui)	20	30	35	45
N(on)	100	80	105	85

1. D'après OzDASL. Les chiffres sont arrondis.

Quelle est la distribution de V ? de H ? On considère la variable H' qui vaut 1 si H =Oui et 0 si H =Non. Calculer $E[H'V]$. On trouve $Cor(H', V) = 0.134$. Quelle corrélation eût-on obtenu si l'on avait choisi le convention inverse pour H' (*Indication :* Exprimer la nouvelle variable H'' en fonction de H').

♣ *Réponse:*

V	0	2	4	8
Proba	0.24	0.22	0.28	0.26

.

H	0	1
Proba	0.74	0.26

La variable obtenue avec la nouvelle convention est $H'' = 1 - H'$. *On a* $Cor(H'', V) = -Cor(H', V) = -0.134$

Régression linéaire

Rappel : $\widehat{a} = \frac{c_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$, $\quad \widehat{b} = \bar{y} - \widehat{a}\bar{x}$, $\quad R^2 = r_{xy}^2$.

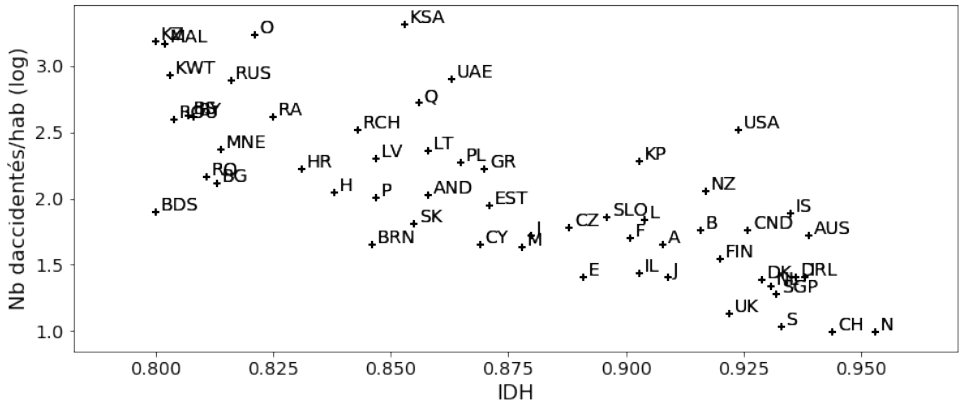
Exercice 12. On considère les données suivantes :

x_i	1	2	3	4	5
y_i	0	0	1	1	2

1. Calculer les moyennes de x et de y. Calculer les variances de x et de y, la covariance de x et y.
2. Donner une équation de la droite de régression de y sur x, la tracer ainsi que les points.

♣ *Réponse:* $\bar{y} = 4/5,\ \bar{x} = 3,\ v_y = 14/25,\ v_x = 2,\ c_{xy} = 1,\ a = 1/2,\ b = -7/10$

Exercice 13. La figure suivante donne pour certain pays le logarithme du nombre de morts sur la route par habitant en fonction de l'indice de développement humain



1. Régression au jugé : Tenter de tracer une bonne droite de régression. Donner son équation.
2. On trouve par le calcul : $m_x = 0.87$, $m_y = 2$, $s_x = 0.047$, $s_y = 0.6$, $c_{xy} = -0.02$. En déduire l'équation de la droite de régression, la corrélation entre les variables, et le R^2.

♣ *Réponse:* $a = -9.05,\ b = 9.87,\ c_{xy} = -0.71,\ R^2 = 0.50$

Exercice 14. On se propose d'étudier l'influence de la température sur la durée d'incubation des œufs de grenouille. On choisit 6 échantillons de 200 œufs chacun. Le nombre y d'éclosions au 22-ème jour est le suivant

Température t_i d'incubation en degrés Celsius	5,8	6,4	6,6	7,4	7,6	8,2
Nombre y_i d'éclosions à la température t_i	131	144	158	171	190	184

1. On donne les valeurs : $\bar{t} = 7,\ \bar{y} = 163,\ s_t^2 = 0.65,\ s_y = 21,\ Cov(t, y) = 16$.
   Calculer le coefficient de corrélation observé et écrire l'équation de la droite de régression de y en t. Qualifier la qualité de l'ajustement.

2. Calculer le nombre d'éclosions prédit pour un échantillon de 200 œufs au 22-ème jour pour une température de 7,5 degrés.

♣ *Réponse:* $y = at + b,\ a = 24.7,\ b = -9.3,\ Cor(y,t) = 0.945,\ R^2 = 0.89,\ N = 175$

Exercice 15. On souhaite exprimer une relation entre le nombre y de téléphones vendus et le budget x investi en publicité. Pour cela, on enquête auprès de 20 compagnies et l'on recueille les données suivantes : $\frac{1}{20}\sum_{i=1}^{20} x_i = 34$ et $\frac{1}{20}\sum_{i=1}^{20} x_i^2 = 1246$, $\bar{y} = 18$, $\overline{y^2} = 340$ et $\overline{yx} = 648$.

1. Calculer l'équation de la droite de régression de y par x.
2. Calculer le coefficient de corrélation, le coefficient de détermination. Qualifier la qualité de l'ajustement.

♣ *Réponse:* $y = ax + b, a = 0.4,\ b = 4.4,\ R^2 = 0.9.$

Exercice 16. On considère le modèle $y_i \simeq ax_i$ à un seul paramètre (au lieu du modèle habituel $y_i \simeq ax_i + b$). Par exemple pour prédire la consommation en litres au cent en fonction de la vitesse. Calculer, en fonction des x_i et des y_i l'estimée de a aux moindres carrés, i.e. celle qui minimise la somme des carrés des erreurs de prédiction de chaque y_i par ax_i.

Exercice 17. (Modèle gravitaire pour les échanges, les migrations, etc.) On suppose que le nombre de personnes de la ville a allant travailler à la ville b suit en gros le modèle idéal $N_{ab} = kd_{ab}^{-\alpha} P_a A_b$ où P_a est a population de la ville a, A_a sa capacité d'accueil (calculé sur la base du nombre de chambres d'hôtels, d'emplois salariés, etc.) et d_{ab} la distance entre les villes. k et α sont des paramètres inconnus. Proposer un modèle de régression linéaire pour des données basées sur n paires villes $\{d_{a_i b_i}, N_{a_i b_i}, P_{a_i}, A_{b_i}, 1 \leqslant i \leqslant n\}$, qui permettra d'estimer les paramètres α et k.
Indication : Il s'agit de choisir y_i et x_i astucieusement.
♣ *Réponse:* $y_i = \ln(z_i) - \ln(P_{a_i}) - \ln(A_{b_i}),\ x_i = (1, -\ln(d_{a_i b_i}))$. *Commentaire : Le mode de calcul des A_b est un truc pas évident. Les $N_{a,b}$ sont assez mal mesurés, ce qui explique aussi l'intérêt des moindres carrés.*

Régression linéaire multiple

Exercice 18. On dispose de deux qualités de papier. Le papier de type 1 a un poids β_1 et le papier de type 2 a un poids β_2 (grammes par feuille). On reçoit n paquets. Le i-ième paquet contient p_i feuilles du type 1 et q_i feuilles du type 2. Chaque paquet est emballé dans une boîte en carton standard, la même quel que soit le nombre de feuilles. On pèse successivement les paquets sur une balance ; le poids mesuré du i-ième paquet est m_i.

1. Proposer une estimation de (β_1, β_2) par régression linéaire. On spécifiera la matrice X.
2. On suppose la boîte de poids nul.
   (a) Que devient le problème ? Résoudre dans le cas $n = 3$, $p_1 = p_2 = q_1 = q_3 = 100$, $q_2 = p_3 = 200$, $y = (1500, 1800, 1600)$ (en gramme).
   (b) Quelles sont les valeurs prédites. Que valent les erreurs de prédiction ?

♣ *Réponse: On a idéalement pour tout i*

$$m_i = \beta_0 + p_i\beta_1 + q_i\beta_2$$

Les erreurs (minimes) de mesures font que l'on ne pourra trouver $(\beta_0, \beta_1, \beta_2)$ qui réalise toutes ces équations simultanément exactement. La méthode est donc de résoudre par moindre carrés, c.-à-d. en minimisant en $\beta = (\beta_0, \beta_1, \beta_2)$

$$SS(\beta) = \sum_i (m_i - \beta_0 - p_i\beta_1 - q_i\beta_2)^2$$

en se plaçant dans le cadre de la régression linéaire multiple (m_i joue le rôle de y_i du cours). La matrice X associée est

$$X = \begin{pmatrix} 1 & p_1 & q_1 \\ \vdots & \vdots & \vdots \\ 1 & p_n & q_n \end{pmatrix}.$$

Commentaire : Si la boîte est de poids nul, on a un pb en dimension 2 sans colonne de 1 (c'est très rare en pratique). Pour la résolution, remarquer que l'on peut tout simplifier par 100. On trouve $\beta = (5, 7)$ (g/feuille) (la formule pour $\widehat{\beta}$ n'est pas à connaître par cœur).

Exercice 19. Soient deux matrices carrées A et B, exprimer l'inverse de AB en fonction de A^{-1} et B^{-1}. On a vu que $\widehat{\beta} = (X^T X)^{-1} X^T y$. Pourquoi ne peut-on pas dire que $\widehat{\beta} = X^{-1} y$?

Exercice 20. Soit un modèle de régression expliquant y en fonction de deux mesures x et z. La matrice X a donc trois colonnes : des 1, x, et z. Soit $R = X^T X$. Expliciter R_{11}, R_{12} et R_{23} en fonction de x et z. On a $R = \begin{pmatrix} 10 & 0 & 0 \\ ? & 17 & 6 \\ ? & ? & 8 \end{pmatrix}$. Compléter la matrice, calculer la covariance entre x et z, calculer s_x^2, calculer R^{-1}.

♣ *Réponse: On note n le nombre de mesures. Explicitons le produit $R = X^T X$:*

$$R = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \\ z_1 & \dots & z_n \end{pmatrix} \begin{pmatrix} 1 & x_1 & z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix} = n \begin{pmatrix} 1 & \overline{x} & \overline{z} \\ \overline{x} & \overline{x^2} & \overline{xz} \\ \overline{z} & \overline{xz} & \overline{z^2} \end{pmatrix}.$$

En particulier $R_{11} = n$, $R_{12} = n\overline{x}$, $R_{13} = n\overline{xz}$. La matrice se complète par symétrie. On a

$$c_{xz} = \overline{xz} - 0 \times \overline{z} = 0.6$$

et

$$s_x^2 = \overline{x^2} - 0 \times 0 = 1.7$$

R est bloc diagonale. Son inverse est formé des inverses des blocs. Comme l'inverse de $\begin{pmatrix} 17 & 6 \\ 6 & 8 \end{pmatrix}$ est $\frac{1}{100}\begin{pmatrix} 8 & -6 \\ -6 & 17 \end{pmatrix}$ on a

$$R^{-1} = \frac{1}{100} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 8 & -6 \\ 0 & -6 & 17 \end{pmatrix}.$$

Exercice 21. Rappeler la formule donnant $\widehat{\beta}$ en fonction de X et y.

1. En déduire que la prédiction $\widehat{y} = X\widehat{\beta}$ s'écrit $\widehat{y} = Py$ pour une certaine matrice P. Exprimer P^2 en fonction de P. Exprimer P^T en fonction de P.
   Indication : Pour le calcul de P^T, on utilisera la formule classique pour la transposée d'un produit de matrices : $(A_1 ... A_n)^T = A_n^T ... A_1^T$. On utilisera également que $(A^T)^T = A$. Par ailleurs, on vérifiera que $X^T X$ est symétrique. On rappelle que l'inverse d'une matrice symétrique est symétrique.
   Note additionnelle : Les relations $P = P^2 = P^T$ indiquent que P est une matrice de projection orthogonale.
   ♣ *Réponse: $\widehat{\beta} = (X^T X)^{-1} X^T y$. Donc $\widehat{y} = X\widehat{\beta} = X(X^T X)^{-1} X^T y$, ce qui donne $P = X(X^T X)^{-1} X^T$. On obtient*

   $$P^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P.$$

   Pour le calcul de P^T, on utilisera la formule classique pour la transposée d'un produit de matrices : $(A_1 ... A_n)^T = A_n^T ... A_1^T$. On utilisera également que $(A^T)^T = A$. Par ailleurs, comme $X^T X$ est symétrique, son inverse S aussi, par conséquent

   $$P^T = (XSX^T)^T = (X^T)^T S^T X^T = XSX^T = P.$$

2. Exprimer le résidu $\widehat{u} = y - \widehat{y}$ en fonction de y et P. Calculer le produit scalaire de $\widehat{y}$ et $\widehat{u}$.
Indication : On utilisera que pour deux vecteurs a et b, identifiés à des matrices colonne, $\langle a, b\rangle = a^T b$.
♣ *Réponse: Comme $\widehat{y} = Py$, on a*

$$\widehat{u} = y - \widehat{y} = (I - P)y.$$

et, en utilisant $P^T = P$ et $P^2 = P$:

$$\langle \widehat{y}, \widehat{u}\rangle = (Py)^T(I - P)y = y^T P(I - P)y = 0.$$

Le vecteur des prédictions est orthogonal au vecteur des résidus.

3. Calculer $X^T\widehat{u}$. Sachant que la première colonne de X ne contient que des 1, que représente la première coordonnée $X^T\widehat{u}$. Que déduire de ces deux constations ?
♣ *Réponse: Comme $\widehat{y} = (I - P)y$, on a*

$$X^T\widehat{u} = X^T(I - X(X^TX)^1X^T)y = (X^T - X^TX(X^TX)^1X^T)y = 0.$$

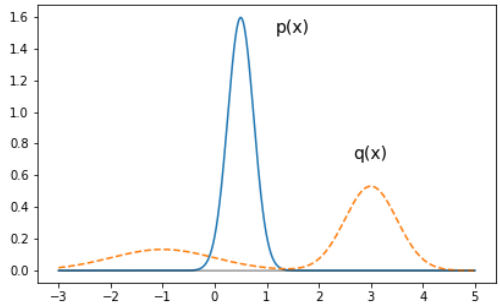
La première coordonnée de $X^T\widehat{u}$ est le produit scalaire de la première ligne de X^T par $\widehat{u}$. Cette ligne ne contenant que des 1, cette coordonnée est la somme des $\widehat{u}_i$. On a donc que la somme des résidus est nulle.

Rappels sur les variables aléatoires

Exercice 22. On sème 100 graines dont chacune a une probabilité $p = \frac{1}{10}$ de germer. On note X le nombre total de graines qui germent.

1. Quelle est la loi de X ?
♣ *Réponse: X est une somme de 100 variables de Bernoulli indépendantes de paramètre 1/10, c'est une* $\mathcal{B}(100, 1/10)$
2. Quelle est son espérance, son écart-type ?

Exercice 23. (On justifiera brièvement les réponses) Soit les deux densités suivantes :



1. Les deux ensembles de données suivants : $A = \{0, 0.3, 0.7, 1\}$ et $B = \{-1, 2, 2.5, 3\}$ représentent chacun des v.a. tirées selon p ou q, mais on ne sait plus lequel a été tiré selon quelle loi. Doit-on attribuer (selon toute vraisemblance) A à p et B à q ou l'inverse ?
♣ *Réponse: La loi de densité $p(x)$ produit des variables assez concentrées autour de 0,5 tandis que $q(x)$ produit des variables de loi plus étalée autour de 3. On attribuera donc A à p et B à q.*
2. L'une des deux lois est gaussienne. Laquelle ?
Indication : S'intéresser à la symétrie.
3. Laquelle a la plus grande variance ?
Indication : Bien comprendre comment l'étalement de la densité est liée à la variance.

4. Laquelle a la plus grande espérance ?
   Indication : On peut considérer que le résultat se voit. Si l'on veut être plus précis, pour calculer l'espérance de la densité en pointillés, définir les surfaces $S_1 = \int_{-\infty}^{1.5} q(x)dx$, $S_2 = \int_{1.5}^{\infty} q(x)dx$ et les intégrales $I_1 = \int_{-\infty}^{1.5} xq(x)dx$ et $I_2 = \int_{1.5}^{\infty} xq(x)dx$, noter que $\int xq(x)dx = S_1\frac{I_1}{S_1} + S_2\frac{I_2}{S_2}$, évaluer les deux quotients à vue d'œil, et considérer que la première bosse fait visiblement, en surface, certainement moins que la moitié de la seconde.

Exercice 24. La durée de vie en années d'une machine à laver suit une distribution $\mathcal{E}(\frac{1}{10})$.

1. Quelle est son espérance de vie ?
   ♣ *Réponse: Commentaire. L'espérance d'une $\mathcal{E}(\lambda)$ doit être connue.*
2. Quelle est la probabilité qu'elle tienne au moins une année ?
   Indication : Pour une v.a. X, de densité p, $P(X > 1) = \int_1^{\infty} p(x)dx$.
3. Quelle est la probabilité qu'elle tienne au moins 10 ans ?
   ♣ *Réponse: 0.37*

Exercice 25. La durée de vie d'une machine à laver suit une loi $\mathcal{E}(1/10)$ (en années). La durée de vie d'un four micro-ondes suit une loi $\mathcal{E}(1/2)$ (en années). Quelle est la probabilité que les deux ustensiles tiennent au moins 2 ans ?
♣ *Réponse: 0.3*

Exercice 26. On suppose qu'il y a deux météos possibles pour un jour donné : soit il pleut, soit il fait beau. On suppose que la météo de chaque jour est indépendante de celle des autres jours et suit la même loi, il y a une probabilité égale à 1/4 de pleuvoir. Soit la variable X_i valant 0 s'il pleut jour i et 1 sinon.

1. Quelle loi suit X_i ?
   ♣ *Réponse:* $\mathcal{B}(1, 3/4)$
2. Quelle est la probabilité qu'il pleuve au moins une fois durant les 5 premiers jours ?
   Indication : Exprimer l'événement complémentaire à l'aide des X_i.
   ♣ *Réponse:* $1 - (3/4)^5$
3. Quelle est la probabilité que durant la première semaine il y ait exactement 2 jours de pluie ?
   Indication : Exprimer cet événement à l'aide des de la somme S des 7 premiers X_i.
   ♣ *Réponse:* $S \sim \mathcal{B}(7, 3/4)$. $P(S = 5) = \binom{5}{7}(1/4)^2(3/4)^5 = ...$
4. Quelle est la probabilité que durant la première semaine il pleuve durant le premier jour et le dernier jour et qu'il fasse beau durant les autres jours ?
   Indication : Exprimer l'événement à l'aide des X_i.
   ♣ *Réponse:* $P(X_1 = X_7 = 0, X_2 = X_3 = X_4 = X_5 = X_6 = 1) = (1/4)^2(3/4)^5$

Exercice 27. Soit X une variable aléatoire de densité $f(x) = \frac{1}{4}x^3 1_{[0,2]}(x) = \frac{1}{4}x^3 1_{0 \leqslant x \leqslant 2}$.

1. Calculer $P(1 \leqslant X \leqslant 2)$.
   ♣ *Réponse:* $P(1 \leqslant X \leqslant 2]) = \int_1^2 f(x)dx = \frac{15}{16}$
2. Que vaut $P(X = 1)$?
3. On tire n nombres indépendamment selon la loi de X.
   (a) Quelle est la probabilité qu'ils soient tous compris entre 1 et 2 ?
   Indication : Si $n = 3$, vous trouvez 0.824.
   (b) Quelle est la distribution du nombre de ceux qui sont compris entre 1 et 2 ?
   ♣ *Réponse:* $\mathcal{B}(n, p)$, *avec* $p = P(1 \leqslant X \leqslant 2) = \frac{15}{16}$
4. Calculer l'espérance et la variance de X.
5. Calculer $E[X^{-1}]$, $E[X^{-3}\cos(X)]$, $E[X^{-2}\cos(X)]$.
   Indication : Les trois résultats sont, par ordre croissant : 0.1, 0.227,2/3

6. Calculer la covariance entre X et X^2.
   Indication : Penser à exploiter les résultats précédents. La réponse est proche de $\frac{1}{3}$.

Exercice 28. Soit X une variable aléatoire de densité $f(x) = cx^2 1_{0\leqslant x\leqslant 1} + \frac{1}{2}1_{2\leqslant x\leqslant 3}$.

1. Que vaut c ?
   Indication : L'intégrale doit faire 1 car c'est une densité.
2. Calculer $P(\frac{1}{4} \leqslant X \leqslant \frac{3}{4})$.
3. Calculer l'espérance et la variance de X. Calculer $E[8X^3 - 60]$.
   Indication : Pour le dernier, penser à la linéarité de l'espérance.

Exercice 29. Soit X une v.a. de densité $cx^{-4}1_{x>1}$. Calculer c, $E[X]$, $Var(X)$.
Indication : Les résultats sont, par ordre croissant : 0.75, 1.5, 3

Exercice 30. On considère une variable aléatoire continue X dont la densité est $f(x) = 3x^2/8$ si $0 \leqslant x \leqslant 2$ et $f(x) = 0$ si $x \notin [0, 2]$.

1. Pourquoi f est-elle bien une densité ?
2. Calculer l'espérance et la variance de X. *Indication :* Leur rapport fait 10.

Exercice 31. Soient X et Y, deux variables aléatoires de Bernoulli de même paramètre p, $0 < p < 1$. On définit les variables aléatoires $S = X + Y$ et $D = X - Y$.

(a) Calculer $\mathrm{Cov}(S, D)$.

♣ *Réponse:* $E[S] = E[X] + E[Y] = 2p$, $E[D] = E[X] - E[Y] = 0$, $E[SD] = E[X^2 - Y^2] = E[X^2] - E[Y^2] = 0$, *où la dernière égalité est vraie parce que X et Y suivent la même loi.*

(b) On suppose X et Y, indépendantes. Les variables S et D sont-elles indépendantes ?
Indication : Calculer $P(S = 0, D = 1)$ et $P(S = 0)P(D = 1)$.

♣ *Réponse:* $P(S = 0, D = 1) = 0$ *car les deux événements $S = 0$ et $D = 1$ sont incompatibles. En revanche,* $P(S = 0) \cdot P(D = 1) = P(X = 0, Y = 0)P(X = 1, Y = 0) = (1 - p^2)(1 - p)p \neq 0$. *On trouve que les variables S et D sont dépendantes.*

Rappels sur la loi normale

Exercice 32. Pour cet exercice, on utilisera la table en fin de document. Soit $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(5, 9)$.

1. Trouver $P(-1 \leqslant X \leqslant 1)$, $P(-2 \leqslant X \leqslant 2)$, $P(0 \leqslant X \leqslant 1)$.
2. Trouver $P(2 \leqslant Y \leqslant 8)$, $P(Y \geqslant 2)$, $P(Y \leqslant 8)$.

Indication : L'ensemble des résultats en ordre croissant est : 0.34, 0.68, 0.68, 0.84, 0.84, 0.96

Exercice 33.

1. Soit X une v.a. à densité de loi $N(0; 1)$. Donner des valeurs approchées de $\mathbb{P}(X > -1)$; $\mathbb{P}(X < -2)$; $\mathbb{P}(1 < X < 2)$ et $\mathbb{P}(|X| < 2)$.
2. Soit Z une v.a. à densité de loi $N(1.75\ ;\ 0.01)$. Donner une valeur approchée de $\mathbb{P}(Z > 1.9)$.
   Indication : La se trouve parmi 0.07, 0.2, 0.13

Exercice 34. On suppose que la taille mesurée en mètre des garçons de 20 ans suit une loi normale de moyenne m et d'écart-type σ. On sait que 84,1% des garçons de 20 ans mesurent moins de 1 m 86 et que 97,7% mesurent plus de 1 m 58. Déterminer m et σ.
♣ *Réponse: On a les équations* $P(\frac{X-m}{\sigma} \leqslant \frac{1.86-m}{\sigma}) = 0.841$ *et* $P(\frac{X-m}{\sigma} \geqslant \frac{1.58-m}{\sigma}) = 0.977$, *d'où* $\frac{1.86-m}{\sigma} = 1$ *et* $\frac{1.58-m}{\sigma} = -2$, *puis* $\sigma = 0.09$ *et* $m = 1.77$.

Exercice 35. Pour un (grand) échantillon d'individus sains, on a étudié la glycémie ; on a constaté que 20% des glycémies sont inférieures à 0.82 g/l et que 32% des glycémies sont supérieures à 0.95 g/l. En supposant que la glycémie suit une loi normale, déterminer la moyenne et l'écart-type de cette loi.
♣ *Réponse: On a les équations* $P(\frac{X-m}{\sigma} \leqslant \frac{0.82-m}{\sigma}) = 0.2$ *et* $P(\frac{X-m}{\sigma} \geqslant \frac{0.95-m}{\sigma}) = 0.32$, *d'où* $P(\frac{X-m}{\sigma} \leqslant \frac{-0.82+m}{\sigma}) = 0.8$ *et* $P(\frac{X-m}{\sigma} \leqslant \frac{0.95-m}{\sigma}) = 0.68$, *puis* $\frac{m-0.82}{\sigma} = 0.84$ *et* $\frac{0.95-m}{\sigma} = 0.47$, *puis* $\sigma = 0.1$, $m = 0.9$.

Exercice 36. 500 personnes ont postulé pour une place, mais 379 ont été refusées parce qu'elles n'étaient pas assez grandes. La taille d'un individu suivant une loi normale de moyenne $m = 171.5$ cm et d'écart-type $\sigma = 5$ cm, estimer la taille minimale exigée.
♣ *Réponse: L'équation* $P(\frac{X-m}{\sigma} \leqslant \frac{t-m}{\sigma}) = 379/500$, *donne* $\frac{t-m}{\sigma} = 0.7$, $t = 175$.

Exercice 37. Soit $(X_k)_{1 \leqslant k \leqslant n}$ des variables indépendantes gaussiennes, $X_k \sim \mathcal{N}(m_k, \sigma_k^2)$. Quelle est la loi de leur somme S ? Quelle est la loi de leur moyenne $M = S/n$? Qu'obtient-on si tous les m_k sont égaux et tous les σ_k aussi ?
Indication (cours) : Une somme de gaussiennes indépendantes est une gaussienne. À vous de trouver espérance et variance.

Exercice 38. (Commencer par l'exercice 37) Le poids des Français, en kg, suit une distribution normale d'espérance 78 et d'écart-type 12.

2. Quelle est la probabilité qu'un homme choisi au hasard pèse plus que 84 kg ?
   ♣ *Réponse: 0.31*
3. Quelle est la probabilité qu'un échantillon de 25 hommes choisis au hasard ait un poids moyen de plus de 84 kg ?
   ♣ *Réponse: 0.006*
4. Quelle est la probabilité qu'un échantillon de 25 hommes choisis au hasard ait un poids moyen entre 74 kg et 82 kg ?
   ♣ *Réponse: 0.9*

Exercice 39. Chaque jour 150 personnes retirent de l'argent à un certain distributeur. Le montant de chaque retrait, en euros, suit une loi $\mathcal{N}(30, 100)$. Combien d'argent doit contenir l'automate en début de journée pour que les clients soient tous servis avec une probabilité supérieure à 0,95 ?
Indication : utiliser l'exercice 37.
♣ *Réponse: 4702*

Loi des grands nombres et théorème-limite central

Exercice 40. On jette n fois un dé et l'on note X_i le résultat du i-ème lancer. On pose $S_n = \sum_{i=1}^n X_i$.

1. Calculer $E[X_1]$ et $\mathrm{Var}(X_1)$.
2. Quelle est la limite de $\frac{1}{n}S_n$ lorsque $n \to \infty$?
3. Quelle est la loi limite de $\frac{1}{\sqrt{n}}(S_n - \frac{7}{2}n)$ lorsque $n \to \infty$?

Exercice 41. Soient $X_1, \ldots X_n$ des variables iid. La variable aléatoire X_1 a pour densité,

$$f(x) = 3x^2 \text{ si } x \in [0,1], \quad 0 \text{ sinon.}$$

1. Quelle est l'espérance de X_1 ? Quelle est la variance de X_1 ?
2. Que donne la loi des grands nombres pour les $X_1, \ldots X_n$?
3. Que donne le TLC pour les $X_1, \ldots X_n$?

Indication : C'est du cours.

Exercice 42. Mêmes questions qu'à l'exercice 41 si la variable aléatoire X_1 est discrète et prend uniformément ses valeurs sur $0, 1, 2, 3, 4$.

Exercice 43. On lance 4096 fois une pièce. Soit S le nombre de pile observés.

1. Quelles sont son espérance m et son écart-type σ ?
   ♣ *Réponse:* $m = 2048$, $\sigma^2 = 4096/4$, $\sigma = 32$.
2. En vous appuyant sur le TLC, proposer une v.a. Z fonction de S dont la loi approche la loi $\mathcal{N}(0, 1)$.
   ♣ *Réponse:* $Z = S/32 - 64$.
3. Pour quelle valeur de c a-t-on $P(-c \leqslant Z \leqslant c) = 0.99$ (on utilisera la table qui se trouve à la fin).
   ♣ *Réponse: 2.58*
4. En déduire un intervalle où S se trouve avec probabilité 0,99.
   Elément de réponse : Sa largeur fait 166.
   ♣ *Réponse:* 2048 ± 83 *soit [1965,2131]*

Exercice 44. On jette un dé 180 fois. On note X la variable aléatoire «Nombre de sorties de 4».

1. Quelle est la loi de X ?
2. En utilisant le théorème-limite central, estimer la probabilité pour que X soit compris entre 29 et 32.
   ♣ *Réponse:* $E[X] = 30$, $Var(X) = 180(1/6)(1 - 1/6) = 25$.
   $P(29 \leqslant X \leqslant 32) \simeq P(29 \leqslant \mathcal{N}(30, 25) \leqslant 32) = P(-1/5 \leqslant \mathcal{N}(0, 1) \leqslant 2/5) = F(0.4) - (1 - F(0.2)) = 0.234$

Exercice 45. Au casino, un joueur roulette parie systématiquement sur le rouge. Il fait mille parties, misant à chaque fois dix euros. Il gagne 10 € si la bille tombe dans une case rouge et perd ses 10 € sinon. Soit X_i la v.a. qui vaut 0 s'il perd à la i-ième partie et 1 sinon.
Rappel : La roulette a 37 cases numérotées de 0 à 36, 18 sont noires, 18 sont rouges, et le 0 est vert.

1. Que dit le théorème-limite central de la somme S des X_i ?
   ♣ *Réponse:* $S \sim \mathcal{B}(1000, p), p = 18/37$. *Son espérance est* $1000p$ *et sa varaince* $1000p(1 - p)$. *La variable centrée-réduite* $Z = (S - 1000p)/\sqrt{1000p(1 - p)}$ *est proche d'une* $\mathcal{N}(0, 1)$.
2. En déduire une valeur approchée de la probabilité qu'il perde moins de deux cent euros. On commencera par exprimer le gain du joueur en fonction de S.
   ♣ *Réponse: Le gain est de 10 si* X_i *vaut 1 et -10 si* $X_i = 0$, *soit* $10(2X_i - 1)$. *Le gain total est donc* $10(2S - 1000)$.
   $P(10(2S - 1000) > -200) = P(S > 490) \simeq P(Z > \sqrt{1000}(490 - 1000p)/\sqrt{p(1 - p)}) = P(Z > 0.222) \simeq 0.22$.

Estimation

Exercice 46. La durée de vie d'une ampoule suit une loi exponentielle $\mathcal{E}(2)$. On mesure la durée de vie de 1000 ampoules (indépendantes). Quelle est la proportion attendue d'ampoules ayant duré au moins 1 an ?
Quel résultat du cours permet de justifier cette approximation ?
Indication : Considérer les variables X_i valant 1 si l'ampoule a tenu au moins 1 an et 0 sinon.
♣ *Réponse: La probabilité qu'une ampoule dure au moins 1 an est de* $\int_1^\infty 2\exp(-2t)\ dt = \exp(-2) \simeq 0,135$. *On peut donc estimer que sur 1000 ampoules, environ 135 dureront au moins 1 an.*
Plus précisément, la variable $S = \sum_{i=1}^{1000} X_i$ *est le nombre d'ampoules ayant tenu au moins 1 an. D'après la loi des grands nombres, la proportion* $\frac{S}{1000}$ *vaut environ* $E[X_i] \simeq 0,135$.

Exercice 47. On tire indépendamment 1000 nombres aléatoires qui sont distribués selon une loi normale $\mathcal{N}(2, 9)$. Combien des nombres tirés, à peu près, sont entre 2 et 3 ? (Utiliser la table des valeurs de Φ.)
♣ *Réponse:* $1000 \times P(2 \leqslant \mathcal{N}(2, 9) \leqslant 3) = 1000 \times P(0 \leqslant \mathcal{N}(0, 1) \leqslant 1/3) = 1000.(0.6293 - 0.5) = 129$

Exercice 48. A la réception de colis, un responsable prélève, au hasard, 25 boîtes qu'il pèse. Soit X_i la masse de i-ème boîte. Il obtient $\sum_{i=1}^{25} X_i = 49,5\,kg$ et $\sum_{i=1}^{25} X_i^2 = 98,3\,kg^2$. On supposera que les masses suivent une loi normale commune $\mathcal{N}(m, \sigma^2)$. Proposer une estimation ponctuelle de m et de σ^2.

Exercice 49. Soient $X_1, \dots X_n$ des variables aléatoires iid de densité $f(y) = \theta^{-1} e^{-y/\theta} 1_{y \geqslant 0}$.

1. Quelle est l'espérance de X_1 ? Proposer un estimateur $\widehat{\theta}$ pour θ.
2. On appelle biais de $\widehat{\theta}$ la quantité $E[\widehat{\theta}] - \theta$. Que vaut-il ?
3. On appelle «erreur quadratique moyenne» (MSE) la quantité $E[(\widehat{\theta} - \theta)^2]$. Que vaut-elle ?

♣ *Réponse:*

1. *Comme $X_1 \sim \mathcal{E}(\theta^{-1})$ on sait que $E[X_1] = \theta$. On peut sinon faire le calcul (avec une IPP, cf le cours), On cherche donc un estimateur de $\mathbb{E}[X_1]$, d'où naturellement*

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i. \quad \textit{(moyenne empirique)}$$

2. *Par linéarité de l'espérance :*

$$\mathbb{E}[\hat{\theta}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \theta = \theta.$$

Ainsi $\mathbb{E}[\hat{\theta}] - \theta = 0$, on dit que l'estimateur $\hat{\theta}$ est sans biais.

3. *Comme $\theta = \mathbb{E}[\widehat{\theta}]$, le MSE est la variance de $\widehat{\theta}$. Il reste donc à calculer $\mathrm{Var}(\widehat{\theta})$. On a*

$$\mathrm{Var}(\widehat{\theta}) = \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_1) = \frac{\mathrm{Var}(X_1)}{n}$$

où l'on a utilisé que les v.a. $X_1, ..., X_n$ sont indépendantes et de même loi. Calculons $\mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$. On a

$$\mathbb{E}[X_1^2] = \int_0^{+\infty} \frac{1}{\theta} y^2 e^{-y/\theta} dy = \left[-y^2 e^{-y/\theta} \right]_0^{+\infty} + 2 \int_0^{+\infty} y e^{-y/\theta} dy = 2\theta^2,$$

et donc $\mathrm{Var}(X_1) = 2\theta^2 - \theta^2 = \theta^2$. On peut ainsi conclure $\mathbb{E}[(\hat{\theta} - \theta)^2] = \frac{\theta^2}{n}$.
Remarque : On aurait pu considérer comme un résultat du cours que la variance de la loi $\mathcal{E}(1/\theta)$ vaut θ^2.

Exercice 50. (Le maximum de vraisemblance) Pour toute variable discrète Y prenant des valeurs y_i avec probabilité $q(y_i)$ (donc $\sum_i q(y_i) = 1$), on appelle $q(Y)$ la vraisemblance de Y, notée génériquement $P(Y)$. C'est la probabilité d'occurrence de l'observation $Y(\omega)$. On notera que pour une suite indépendante $Y_1, \dots Y_n$, $P(Y_1, \dots Y_n) = P(Y_1) \dots P(Y_n)$.

Soit $X_1, \dots X_n$ une suite de v.a.i.i.d. suivant une loi de Bernoulli $\mathcal{B}(1, p)$ (rappel : $p = E[X_i]$).

1. Que représente $p^{X_1}(1-p)^{1-X_1}$? *Indication :* Se souvenir que $X_1(\omega) \in \{0, 1\}$.
2. Utiliser la question précédente pour exprimer la vraisemblance de l'observation $(X_1, \dots X_n)$ en fonction de $S = \sum X_i$.
3. p est en fait inconnu. La vraisemblance de l'observation est vue comme une fonction de p, $p \mapsto f(p)$. Quelle valeur de p réalise le maximum de cette fonction ?
4. On note $\widehat{p}$ la valeur obtenue ; c'est une fonction de $X_1(\omega), \dots X_n(\omega)$. C'est donc une variable aléatoire. Soit $y = n\widehat{p}$. Quelles sont l'espérance et la variance de Y (en fonction de p et n) ? En déduire l'espérance et la variance de $\widehat{p}$?
5. Que vaut le MSE $E[(\widehat{p} - p)^2]$

♣ *Réponse:*

1. *C'est p si $x_1 = 1$ et $1 - p$ sinon, c'est donc la probabilité d'observation de x_1 : $P(X_1 = x_1)$.*
2. $\prod_i p^{X_i}(1-p)^{1-X_i} = \left(\prod_i p^{X_i}\right)\left(\prod_i (1-p)^{1-X_i}\right) = p^S (1-p)^{n-S}$.
3. $\ln(f(p)) = S \ln p + (n - S) \ln(1-p)$. *En dérivant, on voit que le max est en $\widehat{p} = S/n$.*

4. $n\widehat{p} = S \sim \mathcal{B}(n,p)$. $E[n\widehat{p}] = np$, $Var(n\widehat{p}) = np(1-p)$. *Donc* $E[\widehat{p}] = p$, $Var(\widehat{p}) = p(1-p)/n$.
5. $E[(\widehat{p}-p)^2] = Var(\widehat{p}) = p(1-p)/n$

Exercice 51. On veut construire un parc d'attraction. Le projet ne sera considéré rentable que si au moins 1% de la population est intéressé. Dans le but d'estimer la proportion p de personnes qui sont intéressées, on effectue un sondage sur 10000 personnes tirées au hasard dans une population très grande, et $S = 64$ personnes déclarent être intéressées. On appelle X_i la variable aléatoire définie par $X_i = 1$ si la i-ème personne interrogée est intéressée, $X_i = 0$ sinon.

1. Quelle est la loi suivie par chaque X_i ? Quelle est la loi suivie par le nombre S de personnes intéressées par un tel projet dans un tel échantillon de 10000 personnes ?
2. Proposer un estimateur naturel $\widehat{p}$ pour p. Exprimer son espérance et sa variance en fonction de p.
3. En vous appuyant sur le TLC, proposer une v.a. Z fonction de $\widehat{p}$ dont la loi est proche de la loi $\mathcal{N}(0,1)$.
4. En utilisant la table jointe à la fin du document donner un intervalle de confiance centré en $\widehat{p}$, de niveau 0,95, pour la proportion p. Le sondage remet-il en cause la rentabilité du parc ?
   Indication : Trouver c tel que $P(-c \leqslant Z \leqslant c) = 0.95$, puis voir ce que cette identité implique pour p et $\widehat{p}$.
5. Trouver c tel que $P(-Z \leqslant c) = 0.95$. En déduire un nouvel intervalle de confiance pour p.

♣ *Réponse:*

1. X_i *suit la loi de Bernoulli* $\mathcal{B}(p)$. S *s'écrit comme la somme de 10000 v.a. indépendantes de loi* $B(p)$ *donc* $S \sim \mathcal{B}(10000, p)$.
2. *L'estimateur naturel de* $p = E[X_i]$ *est la moyenne empirique* $\widehat{p} = \frac{1}{10000}\sum_{i=1}^{10000} X_i = \frac{S}{10000} = 0.0064$. *On a* $E[\widehat{p}] = p$, *et* $Var(\widehat{p}) = Var(S)/10000^2 = p(1-p)/10000$.
3. *C'est* $\widehat{p}$ *centré-réduit :* $Z = 100(\widehat{p}-p)/\sqrt{p(1-p)}$
4. *Soit* Φ *la fonction de répartition de la loi* $\mathcal{N}(0,1)$. *On a*

$$\mathbb{P}(-c \leqslant Z \leqslant c) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

   On choisit c *de sorte que* $2\Phi(c) - 1 = 0.95$ *grâce à la table de la loi normale. On trouve* $c = 1.96$ *et on a*

$$0.95 = P(-1.96 \leqslant Z \leqslant 1.96) = P\Big(-1.96 \leqslant \frac{100(\widehat{p}-p)}{\sqrt{p(1-p)}} \leqslant 1.96\Big)$$
$$= P\Big(\widehat{p} - \delta \leqslant p \leqslant \widehat{p} + \delta\Big), \qquad \delta = 1.96\frac{\sqrt{p(1-p)}}{100}$$

   On en déduit l'intervalle de confiance de niveau 0.95 pour la proportion p *:* $I = [\widehat{p}-\delta, \widehat{p}+\delta]$. *Comme expliqué dans le cours,* p *étant inconnu, le calcul de* δ *se fait en y remplaçant* p *par* $\widehat{p}$, *ce qui conduit à* $\delta \simeq 0.016$, *puis* $I = [0.0048, 0.008]$. *On n'arrive pas à 1%. Le projet a très peu de chances d'être rentable.*
5. $c = 1,65$. $I = [0, 0.007]$. *C'est encore pire.*

Exercice 52. On effectue un sondage dans le but de déterminer la proportion p d'individus ayant peur en avion. Sur ces 1000 personnes interrogées, seules 253 affirment éprouver la peur de l'avion. Sur la base de ces données, donner un intervalle de confiance à 90% pour p.

Réponse. On doit trouver [0.23 , 0.276].

♣ *Réponse: Comme au 51 :* $Z = \sqrt{\frac{n}{p(1-p)}}(\widehat{p}-p)$. $\widehat{p} = 253/1000$. *Avec pté 0.9 :* $|p-\widehat{p}| \leqslant F(.95)\sqrt{p(1-p)/n} \simeq 1.65\sqrt{\widehat{p}(1-\widehat{p})/n} = 0.023$. $p \in [0.23, 0.276]$.

Exercice 53. Le mathématicien britannique S.G. Soal a mené, dans les années 40, de multiples expériences de télépathie dans lesquelles il demandait à différentes personnes de deviner des cartes. L'une de ces expériences a consisté à faire deviner la valeur d'une carte parmi cinq à Gloria Stewart. L'expérience a été menée 37100 fois. Gloria Stewart a trouvé la bonne valeur 9410 fois.

Considérons l'expérience comme un moyen pour évaluer la probabilité, notée p_*, qu'a Gloria Stewart de deviner la valeur de la carte. On désigne par $\widehat{p}$ la proportion de réponses exactes pour une telle expérience ; c'est une variable aléatoire dont la réalisation est ici $9410/37100 = 0,254$.

1. Si $p_* = 0{,}2$ (Gloria Stewart n'a pas de don particulier), que dit le TCL de la quantité $\widehat{p} - 0.2$?
2. Toujours si $p_* = 0{,}2$, que vaut, pour tout p la probabilité $R(p) = P(\widehat{p} > p)$. On exprimera cette probabilité en utilisant la fonction de répartition $F(.)$ de la gaussienne $\mathcal{N}(0,1)$.
3. La fonction $R(p)$ est la probabilité d'obtenir un score au moins aussi bon que p. Quelle chance aurait eut Gloria Stewart d'obtenir un score aussi bon que $\widehat{p}$ si elle n'avait eu aucun pouvoir extralucide ? On se contentera de majorer ce chiffre minuscule à l'aide de la formule $\mathbb{P}(\mathcal{N}(0,1) > a) \leqslant \frac{1}{a} e^{-\frac{a^2}{2}}$.

Les principes de la statistique mathématique nous inviteraient ici à considérer que madame Stewart est extralucide. Malheureusement, la fraude a été démontrée par la suite [2].

♣ *Réponse:*

1. *Le nombre $\hat{p}$ est une moyenne empirique donc Le TCL nous dit que la loi de $\sqrt{n}\,\dfrac{\hat{p}-\mu}{\sigma}$ converge vers la loi normale centrée réduite. Ici, $\mu = p_* = 0,2$, $\sigma^2 = p_* \times (1-p_*) = 0,2 \times 0,8 = 0,16$ et $n = 37100$. Donc $\sqrt{37100}\,\dfrac{\hat{p}-p_*}{\sqrt{0,16}} = 481,5 \times (\hat{p}-p_*)$ est proche d'une $\mathcal{N}(0,1)$.*
2. *On a*

$$R(p) = P(\hat{p} > p) = P\big(481,5 \times (\hat{p}-p_*) > 481,5 \times (p-p_*)\big) = 1 - F\big(481,5 \times (p-0,2)\big)$$

3. *D'après la majoration indiquée, on a*

$$R(0,254) \simeq P\big(\mathcal{N}(0,1) > 481,5 \times (0,254-0,2)\big) \simeq P\big(\mathcal{N}(0,1) > 26\big) \leqslant \frac{1}{26}\exp(-26^2/2)$$

qui vaut à peu près $6,2\,.\,10^{-149}$. Vu de l'ordre de grandeur du résultat, on peut affirmer qu'il est impossible de deviner autant de cartes par pur hasard.

Exercice 54. Soient $X_1, \dots X_n$ des variables aléatoires i.i.d. On suppose que la variance de X_1 est finie.

1. Donner un estimateur $\widehat{\nu}$ pour l'espérance ν de X_1.
2. Utiliser le TLC pour approcher la loi de $\sqrt{n}\,(\widehat{\nu} - \nu)$.
3. Trouver δ tel que $P(|\widehat{\nu} - \nu| \leqslant \delta) \simeq 95\%$.
4. En déduire un intervalle de confiance asymptotique à 95% pour ν.
5. L'intervalle dépend de σ, qui est inconnu. Donner un estimateur $\widehat{\sigma}^2$ pour la variance σ^2 de X_1. En déduire un nouvel intervalle de confiance. Donner une application numérique en sachant que l'on observe $n = 60$, $\sum_{i=1}^{60} X_i = 49$ et $\sum_{i=1}^{60} X_i^2 = 138$.

♣ *Réponse:* $\nu = \widehat{\nu} \pm \delta$, $\delta = n^{-1/2}\sigma F(.975) = \sqrt{(138/60 - (49/60)^2)/60} \times 1.96 = 0.32$, $\widehat{\nu} = 49/60 = 0.812$

Exercice 55. Il est acquis qu'à la suite d'un premier infarctus du myocarde, 10% des patients font une rechute. La prise d'un anticoagulant pourrait permettre de réduire ce risque. On met en place un essai clinique pour tester l'effet de l'anticoagulant dans la prévention des rechutes. On fait un essai avec $n = 500$ patients. On observe un taux de rechute de $\widehat{\tau} = 8{,}5\%$. Soit τ_* le vrai taux de rechute moyen sous anticoagulant. On veut savoir si l'anticoagulant a un effet, c.-à-d. si $\tau_* < 10\%$.

1. Utiliser le TLC pour approcher la loi de $\sqrt{n}\,(\widehat{\tau} - \tau_*)$.
2. En déduire un nombre q tel que $P(\tau_* - \widehat{\tau} \leqslant q) \simeq 0.95$
3. On considère que l'anticoagulant a un effet significatif au niveau 0,95 si $\widehat{\tau} + q$ (qui majore τ_* avec une certitude à 95%) est inférieur à 10%. Est-ce le cas ?
4. Refaire le calcul en supposant que le même chiffre de 8,5% a été obtenu avec un essai sur 10000 patients.

♣ *Réponse:* $\sqrt{n}\,(\widehat{\tau} - \tau_*) \simeq \mathcal{N}(0, \sigma^2)$, $\sigma = \sqrt{0.085(1-0.085)} = 0.28$, $P(\tau_* - \widehat{\tau} \leqslant q) = P(\mathcal{N}(0,1) \leqslant q\sqrt{n}/\sigma)$ *donc* $q\sqrt{n}/\sigma = 1.65$, $q = 0.02$ *(et* 0.0149 *si* 10000*)*

Exercice 56. Une compagnie d'assurance se propose d'assurer $n = 100\,000$ clients contre le vol. Les sommes en euros (la plupart du temps nulles) $X_1, ..., X_n$ qu'aura à rembourser chaque année la compagnie aux clients

2. https ://www.sceptiques.qc.ca/dictionnaire/soalgoldney.html

sont des v.a. indépendantes d'espérance 75 et d'écart type 750 (p.ex. 7500 fois une $\mathcal{B}(1, \frac{1}{100})$). Quelle prime d'assurance annuelle A la compagnie doit-elle faire payer à chaque client pour que ses frais évalués à 1,5 millions d'euros soient couverts avec une probabilité supérieure ou égale à 0.999 ?
Indication : On exprimera l'événement «les frais sont couverts» en fonction de la prime et de la somme S des X_i, puis on utilisera l'approximation gaussienne.
♣ *Réponse: L'événement est* $S < na - 1.5 \times 10^6$. $P(75n + 750\sqrt{n}\, Z < na - 1.5 \times 10^6) = 0.999$. *D'où* $75n + 750\sqrt{n}\, 3.1 = na - 1.5 \times 10^6$, $a = 15 + 75 + 3.1 \times 750/\sqrt{10^5} = 15 + 75 + 7.3$. *On voit la part du remboursement moyen, celle des frais, et celle de la garantie contre les fluctuations.*

Exercice 57. Un homme politique voudrait évaluer la proportion p de la population française ayant un avis positif sur une certaine question. Il veut un intervalle de largueur 2% contenant p avec probabilité supérieure à 0,999. Proposer une taille d'échantillon suffisante pour répondre à sa commande, en utilisant le théorème-limite central. *Indication* : Le problème est qu'une réponse précise nécessite la connaissance de p ; on contournera cette difficulté en exploitant le fait que $p(1-p) \leqslant \frac{1}{4}$.

Table des quantiles de la loi normale standard

$\Phi(t) = P(X \leqslant t)$ pour $X \sim \mathcal{N}(0,1)$

t	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,5279	0,53188	0,53586
0,1	0,53983	0,5438	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,6293	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,6591	0,66276	0,6664	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,7054	0,70884	0,71226	0,71566	0,71904	0,7224
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,7549
0,7	0,75804	0,76115	0,76424	0,7673	0,77035	0,77337	0,77637	0,77935	0,7823	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,8665	0,86864	0,87076	0,87286	0,87493	0,87698	0,879	0,881	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,9032	0,9049	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,9222	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,9452	0,9463	0,94738	0,94845	0,9495	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,9608	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,9732	0,97381	0,97441	0,975	0,97558	0,97615	0,9767
2	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,9803	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,983	0,98341	0,98382	0,98422	0,98461	0,985	0,98537	0,98574
2,2	0,9861	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,9884	0,9887	0,98899
2,3	0,98928	0,98956	0,98983	0,9901	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,9918	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,9943	0,99446	0,99461	0,99477	0,99492	0,99506	0,9952
2,6	0,99534	0,99547	0,9956	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,9972	0,99728	0,99736
2,8	0,99744	0,99752	0,9976	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861

Table pour les grandes valeurs :

3	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4
0,99865	0,99903	0,99931	0,99952	0,99966	0,99977	0,99984	0,99989	0,99993	0,99995	0,99997