M0SE2014
Probabilités et Statistiques

Philippe Thieullen
Institut de Mathématiques
Université de Bordeaux, CNRS, UMR 5251
F-33405 Talence, France

Philippe.Thieullen@math.u-bordeaux1.fr

Talence, le 30 janvier 2015

L'essentiel du programme

Modalité de contrôle continu

Le cours est composé de 18 séances de 1h20 incluant les travaux dirigés. Il est suivi de 4 séances de 1h20 de travaux pratiques sur machine. Les étudiants peuvent s'associer par binôme ou trinôme. Les TP sont notés et doivent être rendus par courrier électronique en fin de séance. Il est prévu

- 3 contrôles continus de 20-30 mn. Chaque groupe organise son propre contrôle tout en respectant la cadence des séances.
- 1 DS de 1h30. le DS est commun à l'ensemble de l'UE mais chaque enseignant corrige son propre groupe.
- 1 DST de 1h30 aux mêmes conditions que celles du DS.

Evaluation

Contrôles continus	0.2
TP machine	0.2
DS	0.3
DST	0.3

Programme

1. **Statistique descriptive et Indicateurs numériques**
   - Terminologie (population, échantillon $(x_1, x_2, \cdots, x_n)$, taille, caractères, modalités).
   - Notion de caractère statistique (quantitatif, qualitatif, discret, continu), classe (amplitude, milieu).
   - Représentation des données d'un seul caractère : série brute, tableau par valeurs-effectifs (ξ_i, n_i) et par classes-effectifs $([\xi_{i-1}, \xi_i[, n_i)$, par fréquence-effectifs, $f_i = n_i/n$.
   - Diagramme en bâton pour des variables qualitatives, histogramme pour des variables numériques continues (discuter en exercices le cas des classes n'ayant pas toutes la même amplitude), courbe des effectifs cumulés.
   - Moyenne observée : cas d'une série brute, cas d'un échantillon donné par valeurs-effectifs,
   $$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}(n_1\xi_1 + \cdots + n_r\xi_r)$$
   - Médiane $m = q_{50\%}$. Définition dans le cas d'une série brute réordonnée. Formule graphique utilisant la courbe des effectifs cumulés $N_1, \ldots, N_r$ ou des fréquences cumulées, $(F_1, \cdots, F_r)$, dans le cas continu. Formule théorique de la médiane pour un échantillon donné par classes-effectifs
   $$\frac{q_{50\%} - \xi_{i-1}}{\xi_i - \xi_{i-1}} = \frac{50\% - F_{i-1}}{F_i - F_{i-1}} = \frac{\frac{1}{2}n - N_{i-1}}{N_i - N_{i-1}}.$$
   - Variance s_{n-1}^2 ou écart-type observée s_{n-1},
   $$s_{n-1}^2 = \frac{1}{n-1}\Big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\Big).$$
   - Premier quartile $q_{25\%}$, troisième quartile $q_{75\%}$, intervalle interquartile, boxplot (de préférence à boîte à moustache) comme mesure de la dispersion, dans le cas d'une représentation de données par classes-effectifs.

— En TP machine, on verra comment déterminer les quartiles dans le cas d'une série brute de taille n réordonnées par ordre croissant,

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)},$$
$$q_{25\%} = x_{(n/4)}, \quad m = q_{50\%} = x_{(n/2)}, \quad q_{75\%} = x_{(3n/4)}.$$

2. **Espace et mesure de probabilité**

— Notions d'espace fondamental Ω (ou ensemble des épreuves), d'événements $A \subset \Omega$, d'événements élémentaires $\omega \in \Omega$.

— Utiliser des exemples concrets. Construire explicitement $(\Omega, \mathbb{P})$ dans chaque cas (ne pas introduire d'algèbre d'événements !).

— Opérations sur les événements : événement mutuellement incompatibles (ou disjoints $A \cap B = \varnothing$), événement contraire $\bar{A}$ (notation commune imposée), événement certain, impossible.

— Probabilité d'un événement. Probabilité du complémentaire (insister sur son utilisation), de la réunion d'ensembles disjoints (deux ou plusieurs). Formule générale

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

3. **Exercices de révision**

4. **Indépendance et probabilités conditionnelles**

— Indépendance d'événements : définition, exemples.

— Notion d'événements conditionnels. Définition de la probabilité conditionnelle de A sachant B, $\mathbb{P}(A \,|\, B)$ (notation à privilégier sur $\mathbb{P}_B(A)$)l.

— Formule des probabilités composées : $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$.

— Formule des probabiltés totales : système complet d'événements ou partition $\Omega = A_1 \cup \cdots A_r$, formule

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B|A_1) + \cdots + \mathbb{P}(A_r)\mathbb{P}(B|A_r).$$

— Formule de Bayes. (Un exemple type : 15% d'individus d'une certaine population présente une affection A. Un test de dépistage est réalisé. Il s'avère que le test donne 95% de résultats positifs pour les personnes atteintes par A et 10% de résultats positifes pour les personnes non atteintes. Une personne prise au hasard subit le test. Si le test est positif, quelle est la probabilité que cette personne soit atteinte par A ? Si le test est négatif, quelle est la probabilité qu'elle soit indemne ?)

— Présenter plutôt Bayes sous forme d'un tableau ou d'une arborescence.

5. **Variables aléatoires discrètes et lois usuelles**

— Définition générale d'une variable X. Exemple de la loi uniforme, de la loi de Bernoulli. Exemple d'un lancé de dés de la somme des faces de deux dés.

— Définition de la loi de probabilité, de la fonction de répartitions $F(x) = \mathbb{P}(X \leqslant x)$. Faire le lien avec la courbe cumulée.

— Espérance $\mathbb{E}[X]$, variance $\mathrm{Var}(X)$, écart-type σ, espérance d'une fonction de de la variable X, $\mathbb{E}[\phi(X)]$.

— Espérance, variance d'une somme de v.a. Indépendance de deux v.a.

— *Loi de Bernoulli* $\mathcal{B}(p)$. Loi d'une variable X prenant deux valeurs $\{0, 1\}$. Ses paramètres sont donnés par

$$\mathbb{P}(X = 1) = p, \ \mathbb{P}(X = 0) = 1 - p, \quad \mathbb{E}[X] = p, \quad \mathrm{Var}(X) = p(1 - p).$$

— *Loi binomiale* $\mathcal{B}(n,p)$. Loi d'une variable X prenant ses valeurs dans $\{0,1,\dots,n\}$. C'est la loi de la somme de n variables indépendantes et de même loi (i.i.d.) égale à une loi de Bernoulli. Ses paramètres sont donnés par

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad \mathbb{E}[X]=np,\quad \mathrm{Var}(X)=np(1-p).$$

— *Loi de Poisson* $\mathcal{P}(\lambda)$. (Eventuellement en exercice) Loi d'une variable X prenant des valeurs entières, $X=k$, $k=0,1,2,\dots$, quelconques et servant par exemple à modéliser un nombre d'appels téléphoniques par unité de temps.

$$\mathbb{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},\quad \mathbb{E}[X]=\lambda,\quad \mathrm{Var}(X)=\lambda.$$

6. **Exercices de révision**

7. **Variables aléatoires continues et lois uselles**

— Définition au moyen de la notion de densité de probabilité $f(x)$. Exemple de la loi uniforme sur $[0,1]$, sur $[a,b]$.

— Fonction de répartition $F_X(x)=\mathbb{P}(X\leqslant x)$ et quantile d'ordre α, $\mathbb{P}(X\leqslant q_\alpha)=\alpha$ d'une loi à densité.

— Espérance, variance, écart-type. Calculer explicitement ces trois quantités pour la loi uniforme.

— Espérance d'une fonction de la variable X, $\mathbb{E}[\phi(X)]$.

— Cas de plusieurs variables aléatoires. Espérance de la somme, du produit de deux v.a. Cas indépendant : addition des variances.

— *Loi uniforme* $\mathcal{U}(a,b)$. X prend des valeurs dans $[a,b]$ et sa densité est donnée par

$$f(x)=\frac{1}{b-a}\mathbb{1}_{\{a<x<b\}},\quad \mathbb{E}[X]=\frac{b+a}{2},\quad \mathrm{Var}(X)=\frac{(b-a)^2}{12}.$$

— *Loi normale* $\mathcal{N}(\mu,\sigma^2)$ *centrée réduite.* X prend ses valeurs dans $\mathbb{R}$. Ses paramètres sont donnés par

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big),\quad \mathbb{E}[X]=\mu,\quad \mathrm{Var}(X)=\sigma^2.$$

La somme de v.a. normales indépendantes est encore normale.

— *Loi exponentielle* $\mathcal{E}(\theta)$. (Eventuellement en exercice) La variable $X\geqslant 0$ est positive. Ses paramètres sont donnés par

$$f(x)=\theta^{-1}e^{-\frac{x}{\theta}}\mathbb{1}_{\{x>0\}},\qquad \mathbb{E}[X]=\theta,\quad \mathrm{Var}(X)=\theta^2.$$

— *Loi du chi-deux* $\chi^2(n)$ *à* n *ddl.* X prend ses valeurs dans $\mathbb{R}^+$ et a même loi que la v.a. $Z_1^2+\cdots+Z_n^2$. Ses paramètres sont donnés par (ne pas retenir)

$$f(x)=\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)},\quad \mathbb{E}[X]=n,\quad \mathrm{Var}(X)=2n.$$

— *Loi de Student* $\mathcal{T}(n)$ *à* n *ddl.* X prend ses valeurs dans $\mathbb{R}$ et a même loi que la v.a. $X=U/\sqrt{V/n}$ où U et V sont indépendantes, U de loi $\mathcal{N}(0,1)$ et V de loi $\chi^2(n)$. Ses paramètres sont donnés par (ne pas retenir)

$$f(x)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\Big(1+\frac{x^2}{n}\Big)^{-\frac{n+1}{2}},\quad \mathbb{E}[X]=0,\quad \mathrm{Var}(X)=\frac{n}{n-2}.$$

— Exemples d'applications des deux premières lois et utilisation des tables numériques (réduction à la loi normale centrée réduite, exemples de calculs).

— **Premier contrôle continu (20-30 mn)** Contrôle sur l'ensemble des châpitres portant sur les probabilités combinatoires et les variables aléatoires discrètes.

8. **Exercices portant sur les variables aléatoires continues**

9. **Théorème de la limite centrale et applications**

— Epreuves répétées, somme et moyenne

$$Y_n = X_1 + \cdots + X_n, \quad \bar{X} = \frac{1}{n}\Big(X_1 + \cdots + X_n\Big).$$

Espérance de $\bar{X}$, variance de $\bar{X}$. Cas de sommes de v.a. indépendantes

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X_1) = \frac{1}{n}\sigma^2 \quad \text{(cas i.i.d.)}$$

Bien comprendre la différence entre $\mathrm{Var}(10X_1)$ et $\mathrm{Var}(X_1 + X_2 + \cdots + X_{10})$ pour des v.a. iid. Savoir se ramener à la variable centrée réduite

$$Z = \sqrt{n}\Big(\frac{\bar{X} - \mu}{\sigma}\Big) = \sqrt{n}\Big(\frac{\bar{X} - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)}}\Big).$$

— Théorème de la limite centrale,

$$\mathbb{P}\Big(\frac{Y_n - n\mu}{\sigma\sqrt{n}}\Big) = \mathbb{P}\Big(x < \sqrt{n}\frac{\bar{X} - \mu}{\sigma} < y\Big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

ou bien

$$\mathbb{P}\big(n\mu + x\sigma\sqrt{n} < \sum_{i=1}^{n} X_i < n\mu + y\sigma\sqrt{n}\big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

— Approximation d'une loi binomiale, loi de $X_1 + \ldots + X_n$, lorsque les X_i sont des v.a. indépendantes de même loi $\mathcal{B}(p)$, par la loi normale $\mathcal{N}(np, np(1-p))$ lorsque n est grand,

$$\mathbb{P}\Big(x < \frac{X_1 + \ldots + X_n - np}{\sqrt{np(1-p)}} < y\Big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

— Utilisation en exercices des tables de ces lois : tables des fonctions de répartition et tables des quantiles.

10. **Estimation ponctuelle et intervalle de confiance I**

— Statistique inférentielle versus statistique descriptive : réalisation de n v.a. indépendantes $(X_1, \cdots, X_n)$ de loi inconnues $p_\theta(\xi_i)$ dans le cas discret, $p_\theta(x)\, dx$ dans le cas continu

— Définition d'un estimateur ponctuel d'une quantité $\tau(\theta)$: c'est une v.a. $T(X) = T(X_1, \cdots, X_n)$ fonction uniquement de l'échantillon X, sensée représenter τ. Exemple : $\theta = (\mu, \sigma^2)$ et $\tau(\theta) = \sigma^2$ pour une famille de lois normales $\mathcal{N}(\mu, \sigma^2)$.

— Qualité d'un estimateur ponctuel : avec ou sans biais

$$\mathbb{E}_\theta[T(X)] = \tau(\theta), \quad \forall\ \theta.$$

Calculs effectifs d'éstimateurs avec et sans biais par intégration de densité.

— Estimateurs ponctuels classiques.

— estimateur d'une moyenne :

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

— Estimateur d'une proportion ou de la probabilité p d'un événement A :

$$\hat{p} = \frac{1}{n}(\text{nombre de fois que } X_i \text{ réalise } A).$$

— Estimateur d'une variance

$$S_{n-1} = \sqrt{\frac{1}{n-1}\Big((X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2\Big)}.$$

(On admettra que $\mathbb{E}[S_{n-1}^2] = \sigma^2$ et donc que S_{n-1}^2 est un estimateur sans biais de σ^2).

— Par convention, on utilise des lettres majuscules pour les v.a. et des lettres minuscules pour des observations particulières de ces variables. On utilise aussi la convention $\hat{\tau}$,$\hat{p}$,..., pour estimer des quantités τ, p...Mais on a utilisé $\bar{X}$, plus standard, pour estimer μ.

— Définition générale de l'intervalle de confiance d'une quantité τ au risque α ou au seuil de confiance $1-\alpha$:

$$\mathbb{P}_\theta\big(T_{\min}(X) \leqslant \tau(\theta) \leqslant T_{\max}(X)\big) \geqslant 1-\alpha, \quad \forall\ \theta$$

où $T_{\min}(X)$ et $T_{\max}(X)$ sont des estimateurs.

— Intervalle de confiance de la moyenne μ lorsque l'écart-type est inconnu (le cas où l'écart-type σ_0 est connu ne sera pas traité)

$$\mathbb{P}\Big(\bar{X} - t_\alpha\frac{S_{n-1}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t_\alpha\frac{S_{n-1}}{\sqrt{n}}\Big) \geqslant 1-\alpha,$$

où t_α est l'écart d'une loi de Student $\mathcal{T}(n-1)$ à $n-1$ ddl au risque α, soit $t_\alpha = \frac{1}{2}q_{1-\alpha}$ et $\mathbb{P}(|\mathcal{T}(n-1)| > t_\alpha) = \alpha$.

— Intervalle de confiance d'une proportion p

$$\mathbb{P}\Big(\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqslant p \leqslant \hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\Big) \geqslant 1-\alpha,$$

où z_α est tel que $\mathbb{P}(|\mathcal{N}(0,1)| > z_\alpha) = \alpha$. Utilisation des abaques si le temps le permet.

— Il est important de savoir utiliser en exercice les fonctions statistiques des calculettes.

11. **Intervalle de confiance II**

— Intervalle de confiance de la différence des moyennes : Cas de deux échantillons appariés

$$\overline{\Delta X} = \frac{1}{n}(\Delta X_1+\cdots+\Delta X_n), \quad \Delta S_{n-1} = \Big(\frac{1}{n-1}\sum_{i=1}^n(\Delta X_i - \overline{\Delta X})^2\Big)^{1/2}.$$

$$\mathbb{P}\Big(\overline{\Delta X} - t_\alpha\frac{\Delta S_{n-1}}{\sqrt{n}} \leqslant \Delta\mu \leqslant \overline{\Delta X} + t_\alpha\frac{\Delta S_{n-1}}{\sqrt{n}}\Big) \geqslant 1-\alpha,$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n-1)| > t_\alpha) = \alpha$.

— Intervalle de confiance de la différence des moyennes : cas de deux échantillons indépendants

$$\bar{X}_A = \frac{1}{n_A}\Big(X_1+\cdots+X_{n_A}\Big), \quad \bar{X}_B = \frac{1}{n_B}\Big(Y_1+\cdots+Y_{n_B}\Big),$$

$$S_{AB} = \sqrt{\frac{\sum_{i=1}^{n_A}(X_i-\bar{X}_A)^2 + \sum_{i=1}^{n_B}(Y_i-\bar{X}_B)^2}{n_A+n_B-2}}.$$

$$\mathbb{P}\Big(\bar{X}_A - \bar{X}_B - t_\alpha S_{AB}\sqrt{\tfrac{1}{n_A}+\tfrac{1}{n_B}} \leqslant \mu_A - \mu_B \leqslant \bar{X}_A - \bar{X}_B + t_\alpha S_{AB}\sqrt{\tfrac{1}{n_A}+\tfrac{1}{n_B}}\Big) \geqslant 1-\alpha\,,$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n_A+n_B-2)| > t_\alpha) = \alpha$. Dans la pratique, on calcule séparément l'écart-type de chaque échantillon, S_A et S_B, puis l'écart-type global par

$$S_{AB} = \sqrt{\frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{n_A+n_B-2}}.$$

12. **Exercices de révision** Savoir absolument utiliser les fonctions statistiques de la calculette.

13. **Introduction aux tests d'hypothèses I**
    - Hypothèse nulle (principale, préférentielle) H_0, hypothèse complémentaire (alternative, contre hypothèse) H_1, règle de décision, zone de rejet (ou zone critique) $\mathcal{R}$, zone d'acceptation $\mathcal{A}$), variable de décision, valeur critique.
    - Risque de première espèce $\alpha = \mathbb{P}(\mathcal{R}|H_0)$ ou risque de rejeter H_0 à tort, de seconde espèce $\beta = \mathbb{P}(\mathcal{A}|H_1)$.

		en apparence	
		$X \in \mathcal{A}$: on accepte H_0	$X \in \mathcal{R}$: on rejette H_0
en réalité	$\theta \in H_0$ H_0 est vraie	prévision correcte	risque de première espèce
	$\theta \in H_1$ H_1 est vraie	risque de seconde espèce	prévision correcte

    - Un exemple parmi la leçon suivante comme support (par exemple, test d'une proportion ou test de la moyenne) .
    - Retenir la méthodologie d'un test :
      (a) Donner le nom du test.
      (b) Définir l'hypothèse nulle H_0, bilatérale ou unilatérale. Un calcul numérique intermédiaire permet de définir un H_0 plus judicieux.
      (c) Ecrire la zone de rejet $\mathcal{R}$ correspondant à H_0 en faisant apparaître la variable de rejet. Rappeler les définitions des estimateurs entrant dans la définition de $\mathcal{R}$.
      (d) Choisir un seuil de confiance $1-\alpha$ ou niveau d'erreur α et calculer la valeur critique de la variable de décision au vu des données.
      (e) Conclure : accepter ou rejeter H_0 à l'erreur près α. Détailler la réponse sans utiliser le jargon mathématique.
      (f) Calculer la p-valeur du test : c'est-à-dire l'erreur statistique que l'observateur commet en rejetant à tort H_0 compte tenu des données expérimentales.

      Un test d'hypothèse peut donc aussi bien servir de critère de sélection que de preuve de résultat.
    - **Deuxième contrôle continu (20-30 mn)** Le contrôle portera sur les intervalles de confiances et/ou le théorème central limite.

14. **Tests d'hypothèse usuels gaussiens II**
    - Test d'une proportion p (cas des grandes valeurs de n) :

$$H_0 = \{p < p_0\}, \quad H_1 = \{p \geqslant p_0\}, \quad \mathcal{R} = \left\{ \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \geqslant q_{1-\alpha} \right\}$$

$$p_{val} = \alpha \quad \text{tel que} \quad q_{1-\alpha} = z_{2\alpha} = \frac{p_{obs} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

$$H_0 = \{p = p_0\}, \quad H_1 = \{p \neq p_0\}, \quad \mathcal{R} = \left\{ \frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}} \geqslant z_\alpha \right\}$$

$$p_{val} = \alpha \quad \text{tel que} \quad z_\alpha = \frac{|p_{ops} - p_0|}{\sqrt{p_0(1-p_0)/n}}$$

      où $z_\alpha = q_{1-\alpha/2}$ sont les quantiles de la loi normale vérifiant

$$\alpha = 2 - 2 \times \mathbb{P}(\mathcal{N}(0,1) < z_\alpha) = \mathbb{P}(|\mathcal{N}(0,1)| \geqslant z_\alpha) = 1 - \mathbb{P}(\mathcal{N}(0,1) < q_{1-\alpha}).$$

— Test d'une moyenne μ d'écart-type inconnue :

$$H_0=\{\mu<\mu_0\},\quad H_1=\{\mu\geqslant\mu_0\},\quad \mathcal{R}=\left\{\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}}\geqslant q_{1-\alpha}\right\}$$

$$p_{val}=\alpha\quad\text{tel que}\quad q_{1-\alpha}=t_{2\alpha}=\frac{\bar{x}_{obs}-\mu_0}{s^{obs}_{n-1}/\sqrt{n}}$$

$$H_0=\{\mu=\mu_0\},\quad H_1=\{\mu\neq\mu_0\},\quad \mathcal{R}=\left\{\frac{|\bar{X}-\mu_0|}{S_{n-1}/\sqrt{n}}\geqslant t_\alpha\right\}$$

$$p_{val}=\alpha\quad\text{tel que}\quad t_\alpha=\frac{|\bar{x}_{obs}-\mu_0|}{s^{obs}_{n-1}/\sqrt{n}}$$

où $t_\alpha=q_{1-\alpha/2}$ sont les quantiles de la loi de Student $\mathcal{T}_{n-1}$ à $n-1$ degrés de liberté vérifiant

$$\mathbb{P}(|\mathcal{T}(n-1)|\geqslant t_\alpha)=1-\mathbb{P}(\mathcal{T}(n-1)<q_{1-\alpha})=\alpha.$$

— Utiliser les tables statistiques pour calculer z_α et t_α. Remarquer que $z_{2\alpha}=q_{1-\alpha}$ et $t_{2\alpha}=q_{1-\alpha}$ pour les quantiles $q_{1-\alpha}$ respectifs des lois normales et de Student. Ne donner que des encadrements de la p-valeur lorsque la loi est de Student.

— Comparaison de deux moyennes (échantillons appariés) :

$$\overline{\Delta X}=\frac{1}{n}\sum_{i=1}^n\Delta X_i\quad\text{et}\quad \Delta S_{n-1}=\sqrt{\frac{1}{n-1}\sum_{i=1}^n(\Delta X_i-\overline{\Delta X})^2}$$

$$H_0=\{\Delta\mu=0\},\quad H_1=\{\Delta\mu\neq0\},\quad \mathcal{R}=\left\{\frac{|\overline{\Delta X}|}{\Delta S_{n-1}/\sqrt{n}}\geqslant t_\alpha\right\}$$

$$p_{val}=\alpha\quad\text{tel que}\quad t_\alpha=\frac{|\overline{\Delta x}_{obs}|}{\Delta s^{obs}_{n-1}/\sqrt{n}}$$

où $t_\alpha=q_{1-\alpha/2}$ est tel que $\mathbb{P}(|\mathcal{T}(n-1)|\geqslant t_\alpha)=\alpha$.

— Comparaison de deux moyennes (échantillons indépendants) :

$$S_{AB}=\sqrt{\frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{(n_A-1)+(n_B-1)}},$$

$$H_0=\{\mu_A=\mu_B\},\quad H_1=\{\mu_A\neq\mu_B\},\quad \mathcal{R}=\left\{\frac{|\bar{X}_A-\bar{X}_B|}{S_{AB}\sqrt{1/n_A+1/n_B}}\geqslant t_\alpha\right\}$$

$$p_{val}=\alpha\quad\text{tel que}\quad t_\alpha=\frac{|\bar{x}_{A,obs}-\bar{x}_{B,obs}|}{s^{obs}_{AB}\sqrt{1/n_A+1/n_B}}$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n_A+n_B-2)|\geqslant t_\alpha)=\alpha$.

15. **Exercices de révision** Révision portant sur les tests d'hypothèse usuels paramétriques gaussiens (proportion, moyenne, comparaison).

16. **Test du chi-deux d'ajustement**

— Ajustement ou adéquation à une loi discrète à valeurs ou de modalités dans un ensemble fini, $\{\xi_1,\xi_2,\cdots,\xi_r\}$ et de probabilité $(p_1^0,\cdots,p_r^0)$,

$$\mathbb{P}(X=\xi_1)=p_1^0,\quad \mathbb{P}(X=\xi_2)=p_2^0,\quad\cdots\quad \mathbb{P}(X=\xi_r)=p_r^0.$$

— Efectifs théoriques np_j^0, effectifs observés n_j d'estimateur N_j égal au nombre de fois que la variable X_i prend la valeur ξ_j :

$$H_0=\{p_j=p_j^0\},\quad H_1=\{p_j\neq p_j^0\},\quad \mathcal{R}=\{D^2_{r-1}\geqslant q_{1-\alpha}\},$$

$$D^2_{r-1}=\sum_{j=1}^r\frac{(N_j-np_j^0)^2}{np_j^0}\quad\text{et}\quad p_{val}=\alpha\quad\text{tel que}\quad q_{1-\alpha}=\sum_{j=1}^r\frac{(n_{j,obs}-np_j^0)^2}{np_j^0},$$

où D^2_{r-1} suit la loi du chi-deux à $r-1$ ddl et $q_{1-\alpha}$ est tel que $\mathbb{P}(\chi^2_{r-1}\geqslant q_{1-\alpha})=\alpha$.

— **Troisième contrôle continu (30 mn)** Contrôle portant sur les tests usuels gaussiens.

17. **Test du chi-deux d'indépendance**

— Table de contingence de deux v.a. X et Y de modalités $(\xi_1, \cdots, \xi_r)$ et $(\eta_1, \cdots, \eta_s)$, c'est-à-dire une table donnant les effectifs observés du couple $N_{i,j}$ et les effectifs marginaux correspondants : $N_{i*} = N_{i1} + \cdots + N_{is}$ (la notation $*$ désignant une sommation sur l'indice manquant)

	$\boldsymbol{\eta_1}$	$\boldsymbol{\eta_2}$	$\cdots$	$\boldsymbol{\eta_j}$	$\cdots$	$\boldsymbol{\eta_s}$	
$\boldsymbol{\xi_1}$	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1s}	N_{1*}
$\boldsymbol{\xi_2}$	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2s}	N_{2*}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\xi_i}$	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{is}	N_{i*}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\xi_r}$	N_{r1}	N_{r2}	$\cdots$	N_{rj}	$\cdots$	N_{rs}	N_{r*}
	N_{*1}	N_{*2}	$\cdots$	N_{*j}	$\cdots$	N_{*s}	n

— Effectifs observés $n_{j,k}$ du couple (X, Y) d'estimateur $N_{j,k}$ égal au nombre de fois que $X_i = \xi_j$ et $Y_i = \eta_k$. Distribution empirique de (X, Y), $\hat{p}_{j,k} = N_{j,k}/n$, distributions empiriques marginales de X et de Y : $\hat{p}_{j*} = N_{j*}/n$ et $\hat{p}_{*k} = N_{*k}/n$.

— Dans tous les cas, l'hypothèse nulle H_0 est l'indépendance des deux variables. Elle se traduit par $\hat{p}_{j,k} = \hat{p}_{j*}\hat{p}_{*k}$ ou bien par $N_{j,k} = N_{j*}N_{*k}/n$.

— Table des effectifs théoriques au cas où H_0 serait réalisée, c'est-à-dire une table donnant $N_{j*}N_{*k}/n$ dans chaque cas,

	$\eta_{\mathbf{1}}$	$\cdots$	$\eta_{\mathbf{j}}$	$\cdots$	$\eta_{\mathbf{s}}$
$\xi_{\mathbf{1}}$	$N_{1*}N_{*1}/n$	$\cdots$	$N_{1*}N_{*j}/n$	$\cdots$	$N_{1*}N_{*s}/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\xi_{\mathbf{i}}$	$N_{i*}N_{*1}/n$	$\cdots$	$N_{i*}N_{*j}/n$	$\cdots$	$N_{i*}N_{*s}/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\xi_{\mathbf{r}}$	$N_{r*}N_{*1}/n$	$\cdots$	$N_{r*}N_{*j}/n$	$\cdots$	$N_{r*}N_{*s}/n$

— Test du chi-deux d'indépendance :

$$H_0 = \text{« } X \text{ et } Y \text{ sont indépendants »}, \quad \mathcal{R} = \{D^2_{(r-1)(s-1)} \geqslant q_{1-\alpha}\},$$

$$D^2_{(r-1)(s-1)} = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{(N_{j,k} - N_{j*}N_{*k}/n)^2}{N_{j*}N_{*k}/n}$$

$$p_{val} = \alpha \quad \text{tel que} \quad q_{1-\alpha} = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{(n_{j,k} - n_{j*}n_{*k}/n)^2}{n_{j*}n_{*k}/n}$$

$D^2_{(r-1)(s-1)}$ suit la loi du chi-deux à $(r-1)(s-1)$ ddl et le quantile $q_{1-\alpha}$ est tel que

$$\mathbb{P}(\chi^2_{(r-1)(s-1)} \geqslant q_{1-\alpha}) = \alpha.$$

— Numériquement, on donne un encadrement de la p-valeur en se servant des tables des quantiles du chi-deux.

18. **Exercices de révision** Révision portant sur les tests d'hypothèse non paramétriques (indépendance et Chi-deux).

    Il est indispensable de lire l'appendice A sur une introduction au logiciel R avant d'arriver aux séances de TP machine. Le travail en binôme ou en trinôme est possible. Le fichier source contenant les commandes R est rendu en fin de séance par courrier électronique. Les noms des personnes constituant le binôme ou le trinôme sont indiqués en début de fichier. Une personne non présente lors du TP ne peut pas être inscrite.

19. **TP machine I :** Introduction au maniement du logiciel de statistique `R`. Les étudiants peuvent choisir de travailler en binôme ou en trinôme. Les notes de TP machine font partie de la note de contrôle continu. Ce premier TP n'est pas à rendre.
20. **TP machine II :** Modélisation probabiliste. Le TP montre dans une première partie comment créer un échantillon de loi discrète donnée à l'avance. Il montre dans une deuxième partie comment varie la vitesse de convergence dans le théorème central limite pour un échantillon de loi de Bernoulli.
21. **TP machine III :** Statistique inférentielle et intervalles de confiance. Le TP montre dans une première partie comment représenter par des histogrammes et des boxplots des fichiers de données. Il montre dans une deuxième partie comment calculer un intervalle de confiance, d'abords en utilisant les formules du cours, puis en utilisant les fonctions clef-en-main de `R`.
22. **TP machine IV :** Statistique inférentielle et tests d'hypothèse. Le TP montre comment analyser un fichier de données récupérées sur internet contenant le poids de 200 pots de caramel et de chocolat. Il montre d'abords comment réaliser un test de comparaison de moyenne entre les pots de caramel et les pots de chocolat, puis comment réaliser un test d'ajustement des poids des pots de chocolat à une loi normale $\mathcal{N}(\bar{x}, s_{n-1}^2)$.

Statistique descriptive et Indicateurs numériques

Exercice 1. Une enquête sur le groupe sanguin de 36 patients dans un hôpital a donné :

O	A	B	O	A	A	A	O	O
O	A	O	A	B	O	O	O	AB
B	A	A	O	O	A	A	O	AB
O	A	A	B	A	O	A	O	O

Représenter le diagramme en bâtons.

Exercice 2. Le tableau suivant représente la répartition de la population en Europe (hors Union Soviétique de 292 millions) en 1990 en millions d'habitants.

RFA	59	Italie	58	Grande-Bretagne	56
France	56	Espagne	39	Pologne	39
Roumanie	24	Yougoslavie	24	RDA	17
Tchécoslovaquie	16	Pays-bas	15	Hongrie	11
Portugal	11	Belgique	11	Grèce	9
Bulgarie	9	Suède	8	Autriche	8
Suisse	6	Danemark	5	Finlande	5
Norvège	4	Irlande	3	Albanie	3

1. Tracer l'histogramme des tailles des populations selon les classes 0-10, 10-20, 20-30, 30-40, 40-50 et 50-60.
2. Déterminer la courbe des fréquences cumulées.
3. Déterminer la médiane et les deux autres quartiles puis tracer le boxplot.

Exercice 3. On considère un caractère continu pouvant prendre les valeurs suivantes :

classe	7.5-12.5	12.5-17.5	17.5-22.5	22.5-27.5
effectif	3	6	4	1

1. Tracer l'histogramme et la courbe des fréquences cumulées.
2. Déterminer la classe médiane. Déterminer la valeur médiane calculée au prorata du partage de l'effectif de la classe médiane.
3. Déterminer les deux quartiles et tracer le boxplot (ou boîte à moustaches).

Exercice 4. On mesure les qualités de viscosité de certaines huiles de lubrification. Le technicien procède à 150 relevés à raison de 50 relevés par jour sur 3 jours :

viscosité	jour 1	jour 2	jour 3
60	0	1	0
65	2	7	5
70	15	22	38
75	19	18	6
80	11	2	1
85	3	0	0

1. Tracer sur un même graphe les 3 histogrammes.
2. Calculer pour chacune des 3 distributions la moyenne et l'écart-type. Ces résultats reflètent-ils la même tendance qu'en (1) ?
3. Déterminer l'histogramme de l'ensemble de la distribution.
4. Calculer la moyenne et l'écart-type de l'ensemble de la distribution. Comment ces mesures sont-elles reliées à celles trouvées en (2) ?

Exercice 5. Le tableau suivant donne en euros le salaire horaire de 35 employés d'une entreprise :

7	11	7	10	11	9	10	10	12	13
7	8	11	11	14	9	7	9	11	7
9	13	12	14	7	8	7	14	15	9
9	7	11	9	12					

1. Tracer le diagramme en bâton et le diagramme des effectifs cumulés.
2. Calculer la moyenne et la médiane du salaire horaire.

Exercice 6. On compte le nombre de fautes typographiques dans un échantillon de 30 livres d'un certain éditeur :

nb. de fautes	156	158	159	160	162
effectif	6	4	5	6	9

Calculer la moyenne et la médiane du nombre de fautes typographiques par livre.

Exercice 7. On mesure le poids en grammes de 30 nouveau-nés dans une certaine maternité :

2664	2976	3515	3118	3373	3883
2948	3827	3487	3657	2041	3430
2721	3487	3515	3515	3798	2211
3146	3770	3628	2721	3572	3515
3288	3345	3572	2664	3600	3430

1. Tracer la distribution de cette série statistique en utilisant les classes 2000-2800, 2800-3000, 3000-3200, 3200-3400, 3400-3500, 3500-3600, 3600-4000.
2. Calculer la moyenne et l'écart-type.
3. Tracer l'histogramme et la courbe des effectifs cumulés.
4. Déterminer la classe médiane, les deux quartiles et tracer le box -plot.
5. Calculer différemment la médiane ; qu'en pensez-vous ?

Espace et mesure de probabilité

Exercice 8. On considère l'événement A "tirer un As" d'un jeu de 52 cartes. Expliciter l'ensemble fondamental Ω, l'événement A.

Exercice 9. Soit $\Omega = \{a, b, c, d\}$. Déterminer tous les événements contenant à la fois c et d.

Exercice 10. On jette deux dés. Soient A_0, l'événement "la somme des points est paire", A_1, l'événement "la somme des points est impaire" et B, l'événement "la valeur absolue de la différence des points est égale à 4". Combien comptez-vous d'événements élémentaires dans $A_0 \backslash B$, dans $A_1 \backslash B$?

Exercice 11. On jette trois dés non biaisés numérotés de 1 à 3 et identifiables par leur numéro. Chaque dé a 6 faces numérotées de 1 à 6. Déterminez l'ensemble fondamental et calculer dans chacun des cas suivants la probabilité de l'événement

- Trois fois le même chiffre.
- Trois chiffres différents.
- Deux fois le même chiffre et l'autre différent.

Exercice 12. Deux événements A et B d'un espace fondamental Ω ont pour probabilité $\frac{1}{4}$ et $\frac{2}{3}$. La probabilité que les deux événements arrivent simultanément est de $\frac{1}{8}$. Calculer la probabilité que

(1) au moins l'un des deux événements A ou B arrive,

(2) un seul de ces deux événements se produit.

Exercice 13. On suppose que dans un restaurant universitaire on propose deux desserts à chaque repas. La probabilité que l'un des deux soit un yaourt est de 0.4, que l'un des deux soit une orange est de 0.8. La probabilité que les deux soient un yaourt et une orange est de 0.3. Calculer la probabilité que l'on propose :

(a) un yaourt et pas d'orange,

(b) une orange et pas de yaourt,

(c) ni yaourt ni orange.

Exercice 14. On plombe un dé à 6 faces de sorte que la probabilité d'apparition d'une face donnée soit proportionnelle au nombre de points de cette face. On lance le dé deux fois. Quelle est la probabilité d'obtenir une somme des points des deux faces égale à 4 ?

Exercice 15. On dépouille un lot de 100 bulletins sur lesquels figurent les réponses (`oui, non`) à trois questions. Les nombres de reponses `oui` aux questions 1, 2 et 3 sont respectivement 60, 40 et 30 (les candidats peuvent avoir répondu `oui` à d'autres question). Les nombres de bulletins qui ont répondu `oui` aux deux questions 1 et 2, 1 et 3 et 2 et 3 sont respectivement 24, 15 et 12. Enfin, le nombre de bulletins qui ont répondu `oui` aux trois questions est 10. Déterminer les probabilités

(1) d'avoir obtenu deux `oui` et un `non`,

(2) d'avoir obtenu un `oui` et deux `non`,

(3) d'avoir obtenu trois `non`.

Exercice 16. Trois moustiques porteurs du paludisme piquent un homme. Leurs piqures sont indépendantes et chaque moustique a une probabilité de transmettre la maladie à l'homme de 80%. Quelle est la probabilité que l'homme soit atteint de cette maladie ?

Complément : exercices corrigés

Exercice 17. On jette successivement trois dés $D1$, $D2$, $D3$ dont les faces sont numérotées de 1 à 6. On suppose les faces équiprobables pour chaque dé. On appelle "paire", deux faces identiques et la troisième différente. On appelle "brelan", trois faces identiques.

(a) Décrire l'espace Ω de tous les événements. Donner son cardinal.

(b) Quelle est la probabilité d'avoir un brelan ?

(c) Quelle est la probabilité d'avoir une paire ?

(d) Quelle est la probabilité d'avoir trois faces toutes différentes ?

Exercice 18. Un Tour Operator parisien propose à des touristes japonais une formule standard de visites guidées à laquelle ils peuvent ajouter 3 visites supplémentaires et facultatives S1, S2 et S3. Dans S1, on propose la visite des Catacombes, dans S2, la visite de la bibliothèque Beaubourg et dans S3, le château de Versaille. Une étude statistique montre que sur 100 touristes

(a) S1 et S2 sont choisies 17 fois
(b) S1 et S3 sont choisies 16 fois
(c) S2 et S3 sont choisies 26 fois
(d) S1, S2 et S3 sont choisies 11 fois
(e) S1 est choisie 30 fois
(f) S2 est choisie 42 fois
(g) S3 est choisie 71 fois

(1) Sur 100 touristes en moyenne, quel est le nombre d'entre eux choisissant uniquement S1 et S2 ?

(2) Quelle est la probabilité qu'un touriste japonais ne choisissent aucune visite supplémentaires ?

(3) Quelle est la probabilité de ceux choisissant exactement 2 visites ?

Indépendance et probabilités conditionnelles

Exercice 19.

(1) Est-ce que l'assertion suivante est vrai : si A et B sont disjoints, alors $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$?

(2) Soient les événements : H_1 = « gagner le gros lot », $H_2 = \bar{H}_1$ et B = « passer un mois de vacances à Hawai ». On suppose que

$$\mathbb{P}(H_1) = 0.000001, \quad \mathbb{P}(B|H_1) = 0.99 \quad \text{et} \quad \mathbb{P}(B|H_2) = 0.0001.$$

Calculer la probabilité de B.

Exercice 20. On cherche à prévoir le taux d'abstention d'une élection régionale. On admet qu'on peut regrouper la population en 4 groupes, notés A, B, C, D, suivant une proportion respective de 25%, 14%, 39%, 22%. On connait par ailleurs pour chacun des groupes, la probabilité qu'un électeur ne se rende pas aux urnes : soit respectivement pour A, B, C, D, 47%, 29%, 18%, 5%.

Déterminez le taux d'abstention.

Exercice 21. On dispose de deux urnes U_1 et U_2. Dans U_1, il y a 3 boules rouges et 5 boules blanches. Dans U_2, il y a 5 boules rouges et 6 boules blanches. Toutes les boules sont indiscernables en dehors de la couleur.

On considère l'épreuve aléatoire suivante : on tire au hasard une boule de U_1 pour la remettre dans U_2 ; on tire ensuite une boule de U_2 pour la remettre dans U_1. Soient R_1, l'événement "on a tiré une boule rouge de U_1" et R_2, l'événement "on a tiré une boule rouge de U_2".

(a) Calculer $\mathbb{P}(R_1)$.

(b) Calculer la probabilité d'avoir tiré deux boules rouges consécutives. Que peut-on dire de la composition finale des deux urnes ?

(c) Calculer $\mathbb{P}(R_2)$.

(d) Calculer la probabilité qu'à la fin de l'épreuve, la composition de l'urne U_1 n'ait pas changé.

Exercice 22. Une usine fabrique des ampoules électriques à l'aide de trois machines, A, B et C. La machine A assure 20% de la production et 5% des ampoules fabriquées par A sont défectueuses. La machine B assure 30% de la production et 4% des ampoules fabriquées par B sont défectueuses. La machine C assure 50% de la production et 1% des ampoules fabriquées par C sont défectueuses.

— On choisit au hasard une ampoule. Calculer la probabilité que l'ampoule soit défectueuse et produite par A. Même question avec B et C.

— On choisit au hasard une ampoule et on ne s'intéresse qu'aux ampoules défectueuses. Quelle est la probabilté que l'ampoule provienne de A. Même question avec B, avec C.

Exercice 23. Une secrétaire vient d'égarer une lettre. Elle peut se trouver avec probabilité $1 - p$ sur son bureau cachée sous une pile de dossiers ; elle peut aussi se trouver avec probabilité $p/4$ dans un de ces 4 tiroirs qu'on supposera numérotés de 1 à 4. Après avoir ouvert les 3 premiers tiroirs, la secrétaire constate que la lettre ne s'y touve pas. Quelle est la probabilité qu'elle se trouve dans le quatrième tiroir.

Exercice 24. On considère trois urnes, U_1, U_2 et U_3 contenant respectivement, 2 boules noires, 2 boules blanches et dans la dernière, une blanche et une noire. On choisit au hasard une urne et on tire une boule de cette urne. On demande de calculer la probabilité d'avoir choisi la première urne dans les deux cas :

(1) on sait que l'urne contient au moins une boule noire,

(2) la boule tirée est noire.

Exercice 25. Une élection a lieu au scutin majoritaire à deux tours. Deux candidats A et B sont en compétition. Au premier tour, 40% des voix vont à A et 45% à B, le reste étant constitué d'abstentions. Aucun candidat n'ayant la majorité absolue, un second tour est organisé. Tous les électeurs ayant voté la première fois voteront à nouveau. Un sondage indique par ailleurs que 5% des voix de A se reporteront sur B et que 10% des voix de B iront à A. On estime de plus que les deux-tiers des électeurs n'ayant pas voté au premier tour, voteront à raison de 60% pour A et 40% pour B.

(1) Quelle est la probabilité pour qu'un abstentionniste du premier tour vote pour A, pour B ?

(2) D'après ce sondage, quel candidat a la plus forte probabilité d'être élu ?

Exercice 26. On classe les gérants de portefeuille en deux catégories : les spécialistes et les non spécialistes. Lorsqu'un gérant spécialiste achète une valeur boursière pour son client, la probabilité que le cours de celle-ci monte est de 80% ; dans le cas d'un non spécialiste, cette probabilité ne vaut que 50%. Si on choisit au hasard un gérant dans un annuaire professionnel, la proportion des spécialistes est de 20%.

(1) Sans connaître à quel type de gérant on a affaire, calculer la probabilité que la valeur augmente après avis du gérant.

(2) Recalculer alors la probabilité que le gérant soit un spécialiste lorsqu'on sait que la valeur qu'il a achetée a augmenté.

Exercice 27. L'examen clinique d'un individu montre qu'il est atteint d'une pathologie $\mathcal{P}$ pouvant présenter 3 formes A, B, C qu'on cherche à déterminer. On réalise pour cela des tests supplémentaires spécifiques à chacune de ces formes. Si le test positif, la forme correspondante de la pathologie est certainement présente, sinon, on ne peut rien affirmer. On sait que, lorsque cette pathologie est présente, les trois formes A, B, C apparaissent chez les individus dans des proportions 50%, 30% et 20%. On sait aussi que 70% des sujets atteints de la forme A sont détectés par le test correspondant, que 90% atteints de B sont détectés et que 10% atteints de C sont détectés.

(1) Quelle est la probabilité qu'un individu atteint de la pathologie $\mathcal{P}$ présente la forme A de la maladie ?

(2) La forme A étant plus présente, on commence par réaliser le test correspondant. Le test est négatif. Quelle est maintenant la probabilité que cet individu présente la forme A.

(3) Le test pour A n'ayant rien donné, on réalise les deux autres tests correspondant aux formes B et C. Là encore, les tests sont négatifs. Quelle est finalement la probabilité que cet individu présente la forme A ?

Exercice 28. On considère le jeu de stratégie suivant : un candidat se trouve face à 3 portes numérotées de 1 à 3, un prix a été caché aléatoirement derrière l'une d'entre elles. Le candidat se place toujours devant la porte 1. L'animateur (qui sait où se trouve le prix) ouvre une des portes 2 ou 3 et le candidat constate que le prix n'y est pas. Le candidat a-t-il plus de chance d'obtenir le prix en ouvrant la porte 1 plutôt qu'en ouvrant l'autre porte fermée ? On répondra pour cela aux questions suivantes :

1. Le candidat choisit d'ouvrir au hasard la porte 1 ou l'une des deux portes 2 ou 3 non ouverte. Il se trouve face aux éventualités du type suivant $(\omega_1, \omega_2, \omega_3)$ où ω_1 est le numéro de la porte cachant le prix, ω_2 est le numéro de la porte ouverte par l'animateur et ω_3 est le numéro de la porte (autre que celle déjà ouverte) que le candidat veut ouvrir. Calculer la probabilité de chaque éventualité. Quelle est la probabilité de gagner ?
2. Le candidat choisit la stratégie d'ouvrir systématiquement la porte 1 : quelle est la probabilité de gagner ?
3. Le candidat choisit la stratégie d'ouvrir systématiquement l'une des deux portes 2 ou 3 non ouverte : quelle est la probabilité de gagner ?
4. Conclure.

Complément : exercices corrigés

Exercice 29. On constate à un carrefour très dangereux que 3% des conducteurs de moins de 25 ans et 2% des conducteurs de plus de 25 ans se trouvent à l'origine d'un accident. On sait par ailleurs qu'en moyenne sur tout le réseau routier, un conducteur sur cinq a moins de 25 ans. On notera A l'événement "accident" et B l'événement "conducteur de moins de 25 ans".

(1) Calculer la probabilité qu'il y ait un accident.

(2) On vient d'apprendre qu'un accident s'est produit à ce carrefour. Calculer la probabilité qu'il ait été commis par un conducteur de moins de 25 ans.

Exercice 30. Une entreprise informatique fabrique des CD-Rom. Sur 100 CD-Rom fabriqués, 30 possèdent un défaut de fabrication. Pour éliminer les pièces défaillantes, l'entreprise utilise un appareil automatique de contrôle de qualité. Cet appareil peut se tromper sur la qualité du produit. Lors du contrôle, un CD-Rom défectueux est éliminé 9 fois sur 10, un CD-Rom en état de marche est accepté 8 fois sur 10. On notera les événements :

$$D = \text{« le CD-Rom est défectueux »},$$
$$E = \text{« le CD-Rom est éliminé lors du contrôle »}.$$

(1) Calculer les probabilités $\mathbb{P}(D \text{ et } E)$, $\mathbb{P}(D \text{ et } \bar{E})$.

(2) Calculer la probabilité qu'un CD-Rom soit éliminé lors du contrôle.

(3) L'appareil de contrôle élimine un CD-Rom. Quelle est la probabilité qu'il soit défectueux ?

Exercice 31. Un conseil composé d'un président et de 6 membres est amené à se prononcer par oui ou par non sur une motion présentée par le président. Les 7 personnes votent, le président vote oui et la décision est prise à la majorité des voix. Aucun des 6 membres n'est vraiment enthousiaste, chacun a 4 chances sur 10 de voter oui et les 6 membres votent de manière indépendante. Calculer la probabilité que la motion soit acceptée.

Exercice 32. Une entreprise produit un lot de 100 montres par mois. Les lots des trois premiers mois de l'année 2004 étaient de mauvaise qualité et 40% des montres étaient défectueuses. L'entreprise s'en est rendu compte et a amélioré son contrôle de qualité : 10% des montres produites lors des neuf derniers mois étaient défectueuses. On livre début 2005, au gérant d'une bijouterie d'une grande distribution, un lot de montres de cette entreprise. Il aimerait connaître si ce lot provient des trois premiers mois.

(1) Avant même de déballer, calculer la probablité que ce lot provienne des 3 premiers mois.

(2) Le gérant déballe son lot, prend une montre est constate qu'elle est défectueuse. Calculer la probabilité que la montre provienne d'un lot produit au cours des trois premiers mois.

Variables aléatoires discrètes et lois usuelles

Exercice 33. Dans le tableau suivant, la première ligne correspond aux valeurs possibles que peut prendre une variable aléatoire X, la deuxième ligne aux probabilités avec lesquelles ces valeurs apparaissent :

k	-2	0	1	2
$\mathbb{P}(X=k)$	0.1	0.5	0.2	0.2

(1) Calculer l'espérance, la variance, l'écart-type de X.

(2) Déterminer la loi de X^2 de $|X|$. Calculer de deux manières $\mathbb{E}(X^2)$, $\mathbb{E}(|X|)$.

Exercice 34. Une v.a. X peut prendre l'une des trois valeurs 0, 1 ou 2 avec des probabilités positives ou nulles. Déterminer la loi de probabilité de X sachant que $\mathbb{E}[X]=\frac{3}{2}$ et $\mathrm{Var}(X)=\frac{1}{4}$.

Exercice 35. On lance une pièce en l'air. Si "face" arrive, on note dans une variable X le valeur d'un dé qu'on lance au hasard. Si "pile" arrive, on note la somme des valeurs de deux dés lancés au hasard (on supposera qu'ils sont discernables).

(1) Calculer $\sum_{i=1}^6 i = 1+2+3+4+5+6$, $\sum_{i=1}^6\sum_{j=1}^6 i+j = (1+1)+(1+2)+\cdots+(6+6)$ puis $\sum_{i=1}^6 i^2$, $\sum_{i=1}^6\sum_{j=1}^6 (i+j)^2$.

(2) Calculer l'espérance, la variance et l'écart-type de X.

Exercice 36. Une urne contient six boules dont 4 blanches et 2 noires. On extrait une boule de l'urne, on note sa couleur, puis on la remet dans l'urne. On effectue ensuite des tirages sans remise jusqu'à l'obtention d'une boule de la même couleur que précédement. Déterminer la loi de probabilité du nombre X de tirage après remise de la boule tirée initialement.

Exercice 37. Un joueur paie une mise de 150 euros pour participer à un jeu où il doit obtenir deux fois "pile" successivement en quatre lancers au plus d'une pièce de monnaie. Soit X le nombre de lancers pour obtenir deux fois "piles" successsivement ; on prendra $X=+\infty$ s'il n'arrive pas à obtenir deux fois "pile". Son gain est alors $G=2*17^{4-X}-150$ ou bien $G=-150$ si $X=+\infty$. Calculer alors l'espérance de son gain.

Exercice 38. (Difficile) Un garagiste commande au constructeur N voitures. On appelle X le nombre de voitures qu'il pourrait vendre dans l'année. On admet que X suit une loi uniforme sur $\{0,1,2,\cdots,n\}$ pour $n=50>N$. Toute voiture vendue rapporte au garagiste un bénéfice de $a=10000$ euros et toute voiture invendue entraîne une perte de $b=5000$ euros. On appelle $G=$ (bénéfice - perte), le gain du garagiste en fin d'année.

1. Déterminer le gain comme une fonction de X, a et b sur chaque événement $\{X\leqslant N\}$ et $\{X>N\}$.
2. Calculer l'espérance du gain du garagiste en fonction de N, a et b. (On utilisera la formule théorique $\sum_{i=1}^N i = 1+2+\cdots+N=\frac{1}{2}N(N+1)$).
3. Déterminer la valeur numérique de N pour que la commande soit optimale.

Exercice 39. Un lot de bulbes de tulipes a un pouvoir germinatif de 80%. Chaque bulbe contient un et un seul des trois gènes R (rouge), B (blanc) et J (jaune) qui détermine la couleur de la fleur. On suppose que la probabilité que le bulbe possède le gène R, B ou J est égale respectivement à 0.5, 0.1 et 0.4. Dans la suite, on plante 5 bulbes.

1. Quelle est la probabilité d'obtenir 5 fleurs ?
2. Quelle est la probabilité d'obtenir au moins 3 fleurs ?
3. Quelle est la probabilité qu'un bulbe planté produise une fleur rouge ?
4. Soit X la v.a. donnant le nombre de fleurs rouges obtenues. Déterminer la loi de X, son espérance, sa variance et son écart-type.

Exercice 40. La v.a. X représente le chiffre obtenu après le lancer d'un dé à 6 faces.

1. Déterminer la loi de $Y=X(7-X)$. Calculer $\mathbb{E}[Y]$, $\mathrm{Var}(Y)$.

2. (Difficile) On considère maintenant n lancers in épendants et Y_1, Y_2, ..., Y_n, les résultats correspondants. Déterminer la loi de la v.a. égale à la plus grande de ces valeurs. (Indication : déterminer plutôt la fonction de répartition $F(k)$).

Exercice 41. Une centrale de réservations reçoit en moyenne entre 10h et 12h, 72 appels téléphoniques à l'heure. On modélise ce phénomène par une variable aléatoire de Poisson.

1. Déterminer la probabilté pour qu'entre 11h et 11h01 on ait exactement un appel.
2. Déterminer la probabilité pour qu'entre 11h et 11h01 on ait au moins deux appels.

Exercice 42. On estime, qu'en moyenne, 5 bits sur 10000 sont erronés lors de la transmission d'un message entre deux ordinateurs. Sur un message de 75 bits, calculer la probabilité pour qu'il y ait 0, 1, 2, 3 bits erronés. Reprendre ce calcul en prenant la loi de Poisson.

Exercice 43. Au contact d'une parcelle de maïs génétiquement modifié, une étude sur une parcelle saine montre qu'un épi de maïs sur 100 est modifié. Calculer la probabililité que, sur 100 épis, l'un d'entre eux au moins ait été modifié.

Exercice 44. On admet qu'un quart des personnes d'une population de taille n est vacciné contre la grippe. Au cours d'une période d'épidémie, on constate que 15% des personnes malades sont vaccinés et que 8% des personnes vaccinées sont malades.

1. Calculer la probabilité qu'une personne non vaccinée tombe malade.
2. On note S le nombre de personnes malades pendant l'épidémie. Calculer la loi de S (en fonction de n) puis sa moyenne et sa variance.
3. Sur un groupe de $n = 20$ personnes, calculer la probabilité qu'au moins 3 personnes soient malades.

Exercice 45. Une étude réalisée en 2006 par l'institut de sondage CSA montre que 54% des Français sont chrétiens, 31% sans-religion, 4% musulmans, 1% juifs et 10% autres. La France compte alors 60 millions d'habitants. On constitue un échantillon de 7 personnes.

— Quelle est la probabilité d'avoir une majorité de chrétiens dans l'échantillon ?

— Quelle est la probabilité de ne pas avoir de musulmans ?

— Quel est le nombre moyen de sans-religion ?

Exercice 46. On cherche à simuler une variable discrète X prenant les valeurs $1, 2, \ldots, r$ avec probabilité $p_1, p_2, \ldots, p_r$ qu'on suppose strictement positive. Pour cela, on tire au hasard, et de manière uniforme, un nombre réel $U \in]0, 1[$ (le logiciel R le fait très bien par exemple) ; puis on choisit l'unique indice i de $1, \ldots, r$ qui vérifie

$$p_1 + \ldots + p_{i-1} \leqslant U < p_1 + \ldots + p_i$$

(avec la convention, si $i = 1$, $p_1 + \ldots + p_0 = 0$). Montrer alors que X est bien la variable recherchée.

Complément : exercices corrigés

Exercice 47. On choisit au hasard 4 personnes et on admet qu'on a autant de chance de choisir un garçon que de choisir une fille. On appelle X, le nombre de garçons parmi ces 4 personnes et Y, le nombres de couples fille/garçon qu'on peut former.

(i) Donner la loi de X (valeurs prises par X et la probabilité que X prenne ses valeurs).

(ii) Donner numériquement les valeurs de Y en fonction de celles de X. En déduire la loi de Y.

(iii) Calculer l'espérance et la variance de Y.

Exercice 48. L'île Hawaï possède 770 000 habitants dont 60% d'asiatiques, 39% de blancs et 1% de noirs. On tire un échantillon au hasard de 7 personnes.

(1) Quelle est la probabilité d'obtenir une majorité d'asiatiques dans l'échantillon ?

(2) Quelle est la probabilité de n'obtenir aucun noir dans l'échantillon ?

Exercice 49 (♦)**.** Un petite entreprise du bâtiment possède cinq ouvriers travaillant chacun 40 heures par semaine. Comme les affaires marchent mieux, par manque de personnels, le patron de l'entreprise est obligé de refuser du travail et se demande s'il est plus intéressant d'embaucher un nouvel ouvrier que de faire travailler ses ouvriers en heures supplémentaires. Un ouvrier à temps plein coûte à l'entreprise 20 euros de l'heure. Les statistiques du bâtiment montre que l'entreprise pourrait espérer un volume hebdomadaire X en heure de travail réparti selon la distribution suivante :

X	180	190	200	210
Fréquence	0.03	0.09	0.12	0.15
X	220	230	240	250
Fréquence	0.22	0.21	0.13	0.05

Enfin, chaque heure de travail rapporte à l'entreprise 30 euros.

(1) Calculer le bénéfice hebdomadaire moyen (recettes - dépenses) de l'entreprise.

(2) L'entreprise décide d'embaucher un sixième ouvrier. Calculer le bénéfice hebdomadaire moyen. Cette opération est-elle rentable ?

(3) L'entreprise préfère ne pas embaucher de nouveau personnel et chaque ouvrier accepte d'effectuer jusquà 4 heures de travail supplémentaire par semaine au tarif de 25 euros de l'heure. Est-ce rentable ?

(2) Même question que dans (3) en remplaçant 4 heures par 6 heures.

Variables aléatoires continues et lois usuelles

Exercice 50. On considère une variable X à valeurs dans $[0,1]$ définie par la densité $f(x) = kx$.

— Déterminer k pour que f soit une densité de probabilité.

— Calculer $\mathbb{E}(X)$, $\mathrm{Var}(X)$ et $\sigma(X)$.

Même question avec la densité $f(x) = k(1 - |x|)$ définie sur $[-1, +1]$.

Exercice 51. Une v.a. suit une loi de densité $f(x)$ définie par

$$f(x) = \frac{k}{x^3} \quad \text{si } x \geqslant x_0, \quad f(x) = 0 \quad \text{sinon}$$

où $x_0 > 0$ est une constante positive. Déterminer la constante k pour que $f(x)$ soit bien une densité et préciser la fonction de répartition.

Exercice 52. On considère la v.a. X de densité

$$f(x) = \frac{2x}{\theta^2} \quad \text{si } 0 \leqslant x \leqslant \theta, f(x) = 0 \quad \text{sinon}$$

où θ est un nombre positif donné. Déterminer la fonction de répartition $F(x)$, puis calculer $\mathbb{E}(X)$ et $\mathrm{Var}(X)$.

Exercice 53. Soit X une v.a. de densité

$$f(x) = e^{-(x-\theta)} \quad \text{si } x > \theta, \quad f(x) = 0 \quad \text{sinon,}$$

où θ est un paramètre réel donné.

(1) Déterminer la fonction de répartition.

(2) Soit $X_1, \cdots, X_n$ des v.a. indépendantes et de même loi que X et posons $M_n = \min(X_1, \cdots, X_n)$. Déterminer la fonction de répartition puis la densité de la v.a. M_n.

Exercice 54. Soit X une v.a. normale $\mathcal{N}(0,1)$. Calculer (a) $\mathbb{P}(0,53 < X < 2,73)$, (b) $\mathbb{P}(-0,53 < X < 2,73)$, (c) $\mathbb{P}(|X| > 2,27)$, (d) $\mathbb{P}(|X| < 1,45)$. Déterminer les écarts x tels que (e) $\mathbb{P}(|X| > x) = 0.78$, (f) $\mathbb{P}(|X| < x) = 0,22$, (g) $\mathbb{P}(X < x) = 0.93$, (h) $\mathbb{P}(X < x) = 0.22$.

Exercice 55. Sur 100 observations d'une variable X qui suit une loi normale $\mathcal{N}(\mu, \sigma)$, on observe le décompte suivant : $X < 18$, 25 fois, $18 < X < 25$, 40 fois, $X > 25$, 35 fois. Calculer la moyenne de X.

Exercice 56. Une usine fabrique un lot de résistances électriques de 10 ohms. La résistance R de chaque composant est en fait une v.a. distribuée uniformément entre 9.9 et 10.1 ohms. On appelle conductance, l'inverse de la résistance : $C = 1/R$.

1. Déterminer la fonction de répartition $F(x)$ de la variable R ; puis tracer son graphe.
2. Déterminer la fonction de répartition de la v.a. $C = 1/R$; puis déterminer sa densité.

Exercice 57. Une forêt contient 30% de platanes et 70% d'érables. La vitesse de croissance des platanes est une v.a. normale d'espérance 7.8 cm/an et d'écart-type 0.8 cm/an ; celle des érables est une v.a. normale d'espérance 8.2 cm/an et d'écart-type 0.8 cm/an. On appellera V, la vitesse de croissance d'un arbre quelconque tiré au hasard de cette forêt.

1. Calculer la probabilité que V soit inférieure à 7.4 cm/an.
2. La vitesse de croissance d'un arbre a été trouvée inférieure à 7.4 cm/an ; quelle est la probabilité pour qu'il s'agisse d'un platane ?
3. Déterminer la densité de la v.a. V. Déterminer son espérance, puis son écart-type.

Exercice 58. On cherche à simuler une variable aléatoire X de loi exponentielle et de paramètre $\theta > 0$. On rappelle que la fonction de répartition de cette loi est donnée par $F_X(t) = \mathbb{P}(X \leqslant t) = 1 - \exp(-\theta t)$. Pour cela, on tire au hasard, et de manière uniforme, un nombre réel $U \in]0,1[$; puis on pose $X = \frac{1}{\theta} \ln \frac{1}{U}$. Montrer alors que X est bien la variable recherchée.

Complément : exercices corrigés

Exercice 59. On considère une variable aléatoire continue de densité f donnée par :

$$f(x) = kx^2 \quad \text{pour } x \in [0,1] \text{ et } \quad f(x) = 0 \quad \text{partout ailleurs,}$$

où k est un facteur de normalisation.

(1) Déterminer k.

(2) Calculer la fonction de répartition $F(x) = \mathbb{P}(X \leqslant x)$ pour tout $x \in \mathbb{R}$. Puis dessiner grossièrement le graphe de F.

(3) Calculer $\mathbb{E}(X)$, $\mathrm{Var}(X)$ et l'écart type de X.

Exercice 60. Un composant électronique ne fonctionne que si sa tension est comprise entre 22 Volts et 26 Volts. Son alimentation est une v.a. qui suit une loi normale de moyenne 24 Volts et d'écart type α Volts où α est un paramètre qu'on ajustera.

(1) Dans cette question, on suppose que $\alpha = 1,8$.

(a) Calculer la probabilité que le composant fonctionne.

(b) On suppose que le composant est détruit définitivement si sa tension dépasse 29 Volts. Calculer alors la probabilité qu'il soit détruit.

(2) Quelle valeur faudrait-il donner à α pour que la probabilité que le composant fonctionne soit au moins 85% ?

Exercice 61. On définit sur l'ensemble $[-1,1]$ une fonction f par

$$f(x) = \begin{cases} b(1+x) & \text{si } -1 \leqslant x \leqslant 0 \\ b(1-x) & \text{si } 0 \leqslant x \leqslant 1 \end{cases}$$

(i) Dessiner le graphe de f.

(ii) Déterminer la valeur de b telle que f soit la densité d'une variable aléatoire définie sur $[-1,1]$. Soit Z une telle v.a.

(iii) Calculer $\mathbb{E}(Z)$.

(iv) Calculer $\mathrm{Var}(Z)$ et $\sigma(Z)$.

Exercice 62. On considère une variable aléatoire X numérique de densité $f(x)$ donnée par

$$\begin{cases} f(x) = k(3+x) & \text{si } -3 \leqslant x \leqslant 0 \\ f(x) = k(3-x) & \text{si } 0 \leqslant x \leqslant 3 \end{cases} \qquad f(x) = 0, \quad \text{si } x < -3 \text{ ou } x > 3.$$

(1) Tracer l'allure du graphe de f. Déterminer la valeur de k pour que f soit bien la densité d'une variable aléatoire.

(2) Déterminer la fonction de répartition et tracer l'allure de son graphe.

(3) Déterminer la probabilité que X dépasse la valeur $x_0 = 2$.

Exercice 63 (⋆). Un automobiliste se rend à son travail 225 jours par an. Il a 2 options. Dans la première option, il achète un abonnement d'une place de parking à l'année pour un prix de S euros. Dans la seconde, il décide de stationner chaque fois sur un emplacement interdit et risque alors une amende de 10 euros par jour s'il est verbalisé. On ne peut être verbalisé qu'au plus une fois par jour. La probabilité d'être verbalisé est de 20% chaque jour. On appelle X le nombre de fois que l'automobiliste est verbalisé au cours d'une année de travail dans le cas où il choisit la deuxième option.

(1) Déterminer la loi de X. Calculer son espérance et sa variance.

(2) On appelle Y le total des amendes d'une année dans le cas de la deuxième option. Exprimer Y en fonction de X. Calculer $\mu = \mathbb{E}(Y)$ et $\sigma^2 = \mathrm{Var}(Y)$.

On admettra qu'on peut approcher la loi de Y par la loi normale $N(\mu, \sigma)$.

(3) Calculer la probabilité $\mathbb{P}(450 < Y < 458)$.

(4) Si l'automobiliste choisit la deuxième option, on dira qu'il est «gagnant» si, avec une probabilité d'au moins 0.75, le total des amendes est inférieur à S. On suppose ici que $S = 500$ euros. Est-il gagnant ?

(5) Le gérant du parking cherche à attirer cet automobiliste en lui proposant un abonnement promotionnel S de sorte que, s'il choisissait la deuxième option, il obtiendrait $Y > S$ avec une probabilité au moins égale à 0.75. Calculer le prix S de l'abonnement.

Théorème de la limite centrale

Exercice 64. Une entreprise comprend 900 employés. Elle dispose d'une cantine de 500 places qui assure deux services. On suppose que chaque employé choisit indifféremment l'un ou l'autre service. On désigne par X la v.a. qui associe à un jour donné, le nombre de personnes choisissant le premier service.

1. Déterminer la loi de X. Calculer $\mathbb{E}(X)$ et $\mathrm{Var}(X)$.
2. En approchant X par une loi normale, déterminer la probabilité pour qu'un jour donné on refuse des clients à la cantine.

Exercice 65. On admet que 15% des réservations d'un train à réservation obligatoire sont annulées dans l'heure précédant un départ sans pour autant être remises à la vente.

1. Sur 1000 réservations effectuées au moins une heure à l'avance, déterminer la probabilité que plus de 160 d'entre elles soient annulées à la dernière minute.
2. Un train contient exactement 1600 places assises et la SNCF décide d'accepter plus de réservations que de places assises sachant que certaines de ces places seront libérées à la dernière minute. Quelle est le nombre maximum de réservations que la SNCF va accepter pour que la probabilité d'être en sur-réservation (nombre de passagers dans le train supérieur au nombre de places assises), après annulation de dernière minute, soit inférieure à 1% ?

Exercice 66. La durée de vie d'un chauffage électrique est une variable aléatoire T de densité :

$$\begin{cases} f(t) = 0, & \text{si } t < 0 \\ f(t) = \lambda^2 t e^{-\lambda t}, & \text{si } t \geqslant 0 \end{cases}$$

où λ est un réel positif. L'unité de la durée est l'heure.

1. On prend $\lambda = 2.10^{-4}$ h^{-1}. Déterminer $\mathbb{P}(T < 500)$ et $\mathbb{P}(T > 10000)$.
2. On prend toujours $\lambda = 2.10^{-4}$ h^{-1}. Déterminer la probabilité de trouver dans un lot de 600 appareils choisis indépendamment, plus de 6 appareils ayant une durée de vie inférieure à 500 h, plus de 220 appareils ayant une durée de vie supérieure à 10 000 h.
3. On mélange 10 appareils de fabrication caractérisée par $\lambda = a$ et 5 appareils de fabrication caractérisée par $\lambda = b$. On choisit au hasard un appareil de ce mélange. On appelle X sa durée de vie. Déterminer la densité de X. Déterminer son espérance et sa variance.

Exercice 67. On suppose qu'une ampoule électrique est défectueuse avec probabilité 0.2. En utilisant une approximation par la loi normale, calculer la probabilité que 20 ampoules au moins soient défectueuses sur un lot de 160.

Exercice 68. Une tréfilerie fabrique des câbles métalliques conçus pour résister à de lourdes charges. Leur résistance à la rupture est en moyenne égale à 3 tonnes, avec un écart-type égal à 200 kg. Un contrôle de qualité de la fabrication est effectué sur un échantillon de $n = 100$ câbles.

1. Calculer la probabilité que la résistance moyenne des câbles de l'échantillon soit comprise entre 2.98 tonnes et 3.02 tonnes.
2. On estime que la taille n de l'échantillon ne garantit pas une probabilité suffisante. Déterminer la taille minimale pour que le calcul (1) soit sûr à 90%.

Exercice 69. Deux modèles d'ampoules électriques qualitativement différentes sont produits par une même entreprise. Les ampoules de type A ont une durée de vie moyenne de 1200 h avec un écart-type de 200 h, les ampoules de type B ont une durée de vie moyenne de 1000 h avce un écart-type de 100 h. On prélève au hasard un échantillon de 100 ampoules de type A et 150 ampoules de type B.

Quelle est la probabilité que la durée moyenne de vie observée sur l'échantillon des ampoules de type A soit supérieure de 160 h à celle constatée sur l'échantillon des ampoules de type B ?

Exercice 70. Une entreprise fabrique des cigarettes. Le fabriquant garantit que leur masse suit la loi $\mathcal{N}(\mu_0, \sigma_0^2)$ où $\mu_0 = 1.2$ gr et $\sigma_0 = 0.063$ gr. Il se peut que, par suite d'un déréglement de la machine qui les fabrique, la masse des cigarettes suit la loi $\mathcal{N}(\mu, \sigma_0^2)$. On prélève au hasard 30 cigarettes, leur masse moyenne est de 1.25 gr.

1. On pose $\delta = 0.02$ gr. Calculer la probabilité que la moyenne des masses des cigarettes de l'échantillon soit comprise entre $1.2 - \delta$ gr et $1.2 + \delta$ gr dans le cas où la machine est bien réglée.
2. Trouver l'intervalle centré $[1.2 - \delta, 1.2 + \delta]$ pour que la probabilité calculée en (1) soit égale à 98%.
3. A 2% près d'erreur, l'échantillon précédent montre-il que la machine est bien réglée ?

Complément : exercices corrigés

Exercice 71. Un éditeur publie des livres pouvant contenir des pages erronées. On suppose que la probabilité qu'une page ait au moins une erreur est de 5%.

(1) On appelle X le nombre de pages erronées dans un livre de 300 pages. Quelle est la loi que suit X ? Donner la formule $\mathbb{P}(X = k)$ en fonction de k.

(2) Calculer l'espérance et la variance de X.

(3) On approche la loi de X par une loi normale dont on précisera les paramètres. L'éditeur rejette un livre s'il contient plus de 20 pages erronées. Combien de livres, sur 100 livres imprimés, l'éditeur rejette-t-il ?

Exercice 72 (♦). Une cabine téléphérique peut accepter 50 personnes et ne doit pas dépasser en poids 4700 Kg. On suppose que chaque personne est une v.a. suivant une loi normale de moyenne 80 Kg et d'écart type 12 Kg. On suppose aussi que chaque personne a le droit d'emporter avec elle des bagages et que chaque bagage est une v.a. suivant une loi normale de moyenne 10α Kg et d'écart type 5α Kg où α est un paramètre réel positif. On suppose enfin que X et Y sont indépendantes.

(i) On note $W = X + Y$ la variable aléatoire qui représente le poids d'une personne avec ses bagages. Pourquoi W suit-elle une loi normale ? Calculer en fonction de α, $\mu = \mathbb{E}(W)$ et $\sigma = \sigma(W)$. Vérifier que pour $\alpha = 1$ on a bien $\mu = 9O$ kg et $\sigma = 13$ kg.

(ii) On note W_{50} le poids cumulé de 50 personnes avec leurs bagages ; les personnes sont choisies arbitrairement et de façon indépendante. Calculer en fonction de α l'espérance $\mathbb{E}(W_{50})$ et l'écart-type $\sigma(W_{50})$.

(iii) On suppose que $\alpha = 1$. Déterminer la probabilité que la cabine soit en surcharge.

(iv) Quelle valeur faut-il donner à α pour que probabilité que la cabine soit en surcharge soit inférieure à 1%.

Exercice 73. Jean envisage d'être représentant de commerce dans le monde de l'édition. La somme qu'un tel représentant peut gagner par jour, est une v.a. que l'on suppose normale, de moyenne 400 euros et d'écart type 50 euros. Une journée est appelée bonne s'il gagne plus de 422 euros.

(1) Calculer la probabilité qu'une journée soit bonne.

(2) Jean fait un test sur 5 jours. Soit Y le nombre de bonnes journées sur cette période. Quelle est la loi de Y ? Quelle est la probabilité que Jean ait fait au moins 3 bonnes journées ?

(3) Satisfait de son essai, Jean travail 275 jours par ans. Soit W le nombre de bonnes journées réalisées sur cette période.

  (a) Quelle est la loi de W, sa moyenne μ, son écart-type σ (arrondi à la journée) ?

  (b) Par quelle loi peut-on approcher la loi de W ?

  (c) En utilisant cette approximation, calculer la probabilité que Jean réalise entre 75 et 100 bonnes journées dans l'année.

Estimation ponctuelle et intervalle de confiance

Exercice 74. On considère un échantillon de loi gaussienne de paramètre (μ, σ^2). On note $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ et $\bar{Y} = \bar{X}(1 - \bar{X})$.

1. Calculer $\mathbb{E}[\bar{Y}]$. (Indication : $\mathbb{E}[(\bar{X} - \mu)^2] = \mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2$.)
2. Peut-on dire que $\bar{Y}$ est un estimateur sans biais de $\mu(1 - \mu)$?
3. Comment modifier $\bar{Y}$ pour qu'il devienne sans biais ? (Indication : $\mathbb{E}[S^2_{n-1}] = \sigma^2$.)

Exercice 75. On considère un échantillon $(X_1, \ldots, X_n)$ de loi exponentielle de paramètre λ (de densité $\lambda e^{-\lambda x}$, pour $x \geqslant 0$). On cherche à estimer $e^{-\lambda}$. Pour cela on définit un nouvel échantillon $(Y_1, \ldots, Y_n)$ où

$$Y_i = 1 \text{ si } X_i > 1, \quad Y_i = 0 \text{ si } X_i \leqslant 1.$$

On pose $\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$. Montrer que $\bar{Y}$ est un estimateur sans biais de $e^{-\lambda}$.

Exercice 76. Le temps de vie en heure d'un certain composant électronique est supposé distribué suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$. On réunit les données suivantes :

4671	3331	5270	4973	1837
7783	6074	4777	5263	5418

Calculer un intervalle de confiance de la moyennne de la durée de vie de ce composant.

Exercice 77. Dans un échantillon de 197 pommes, on constate que 19 d'entre elles sont abîmées. Déterminer un intervalle de confiance de la proportion de pommes abîmées.

Exercice 78. Un échantillon aléatoire de 16 voitures est soumis à un contrôle de vitesse. On mesure les vitesses suivantes en km/h :

49	71	78	58	83	74	64	86
56	65	55	64	65	72	87	56

Construire un intervalle de confiance de la moyenne des vitesses à 95%.

Exercice 79. Une enquête réalisée par "The Gallup Organization, Hongrie" en 2009, révèle que 60% $\pm$2% des français sont contre l'adhésion de la Turquie dans l'Europe des 27 dans un avenir relativement proche. Le sondage a été conduit sur un échantillon de 914 personnes majeures. Quel niveau de confiance, l'institut de sondage a t-il utilisé ?

Exercice 80. Lors d'un contrôle de fabrication de certaines pièces mécaniques, on constate que sur 150 pièces, 17 sont défectueuses.

1. Déterminer un intervalle de confiance au risque 5% de la proportion de pièces défectueuses.
2. Combien de pièces doit-on contrôler pour que la proportion observée soit correcte à 1% près au risque de 5%. ?

Exercice 81. On considère 10 sujets pris comme leurs propres témoins. On cherche à comparer deux soporifiques A et B administrés à chaque sujet à raison d'un comprimé par nuit et par sujet. Le tableau suivant indique le nombre d'heures de sommeil des 10 sujets, une pemière fois pour le soporifique A et une semaine plus tard pour le soporifique B.

sujet	1	2	3	4	5	6	7	8	9	10
A	6	7	7	8	8	8	8	9	9	10
B	6	6	5	5	7	7	6	7	7	8

Déterminer l'intervalle de confiance au risque de 5% de la différence des résultats moyens obtenu entre A et B.

Exercice 82. On s'intéresse à la différence des espérances de vie entre l'Europe et l'Afrique. Différents sondages, pris sur des échantillons représentatifs dans chaque continent, donnent les résultats suivants :

Afrique		**Europe**	
orientale	51.1	**de l'Est**	68.6
centrale	47.2	**du Nord**	79.0
septentri.	68.9	**du Sud**	79.5
australe	48.9	**de l'Ouest**	80.0
occidentale	50.5	Source : Nations Unis	
sub-sahari.	50.1	2007	

Déterminer l'intervalle de confiance de la différence des espérances de vie entre l'Europe et l'Afrique à 2 % près d'erreur statistique.

Exercice 83. On cherche à calculer l'efficacité d'un nouvel engrais dans la culture de l'asperge. On partage pour cela 5 parcelles d'asperges en deux parties égales. L'une des moitiés de chaque parcelle, choisie aléatoirement, reçoit un nouvel engrais ; l'autre moitié n'est pas traitée. Le rendement par parcelle est le suivant :

parcelle	1	2	3	4	5
sans	327	204	246	312	279
avec	321	216	264	303	291

On demande de construire un intervalle de confiance au seuil 95 % de la différence de rendement. On introduira pour cela un modèle statistique gaussien $\mathcal{N}(\mu, \sigma^2)$ de différence de rendement moyen μ et d'écart type σ à estimer. On définira des estimateurs, puis on construira un intervalle de confiance théorique. On soignera enfin l'application numérique en précisant les valeurs tirées des tables de loi.

Exercice 84. On désire évaluer le nombre N d'individus d'une espèce animale vivant sur une île. On commence pour cela par capturer 800 individus que l'on marque et relache juste après. Après avoir laissé les individus se remélanger dans la polulation globale, on capture à nouveaux 1000 individus parmi lesquels on dénombre 250 individus marqués. En déduire un intervalle de confiance de N à une erreur près de 5%.

Complément : exercices corrigés

Exercice 85. Une enquête a été effectuée sur le prix des lecteurs DVD dans les supermarchés. Pour un modèle précis, les prix suivants (en euros) ont été relevés dans huit supermarchés :

150	175	130	175	180	145	150	135

(1) En admettant que le prix d'un lecteur DVD se comporte comme une variable aléatoire normale X, calculer une estimation $\bar{x}$ de l'espérance mathématique $\mathbb{E}(X)$ et une estimation s de l'écart-type $\sigma(X)$.

(2) Donner les intervalles de confiance à 90% et 95% pour l'espérance mathématique de X.

Exercice 86. Une audimétrie a permis détablir, avec un risque d'erreur de 5%, que l'audience d'une émission télévisée était dans l'intervalle $[35\%, 45\%]$. Afin d'affiner la mesure, on décide de sonder des téléspectateurs. Combien faut-il en sonder au minimum pour que, avec le même risque d'erreur de 5%, et en supposant que l'on obtienne la même estimation ponctuelle de 40% d'audience, on ait un intervalle de confiance de longueur 2% au lien de 10%.

Tests d'hypothèse usuels gaussiens

Exercice 87. Dans un échantillon de 197 pommes, on constate que 29 d'entre elles sont abimées. Au niveau de confiance 95%, voudriez-vous conclure que 10% des pommes sont abimées ?

Exercice 88. Un nouveau vaccin contre un certain virus a été testé sur 147 individus sélectionnés au hasard. On a constaté que 61 d'entre eux étaient encore atteints par la virus. Sans traitement particulier, on sait que le virus atteint une personne sur deux au moins. Voudriez-vous conclure au risque 5% que la vaccin a été efficace ? Quelle est la p-valeur du test ?

Exercice 89. Un sondage est effectué avant une élection. Sur 887 personnes intérogées, 389 votent pour le candidat A et 369 pour le candidat B.

1. Au risque 5%, voudriez-vous conclure que le candidat A recevra plus de 40% de votes ?
2. Au risque 20%, voudriez-vous conclure que le candidat B recevra plus de 40% des votes ?

Exercice 90. On désigne par p la probabilité d'observer un phénotype donné Q sur un individu issu d'un certain croisement. Pour tester l'hypothèse $p = 9/16$ contre $p = 9/15$, on observe 2400 individus. Si le nombre d'individus présentant le phénotype Q est inférieur à 1395, on choisit $p = 9/16$. Si le nombre d'individus est au contraire supérieur à 1395, on choisit $p = 9/15$. Justifier le principe de ce test et calculer son niveau.

Exercice 91. Une entreprise fabrique des cigarettes ayant une masse moyenne de 1.2 g et un écart-type de 0.07 g. On admet que l'écart-type est constant dans le temps mais que la masse des cigarettes peut fluctuer. A un moment donné, on prélève au hasard 30 cigarettes, leur masse moyenne étant égale à 1.24 g. Pensez-vous que la machine est déréglée ? Déterminer la p-valeur du test.

Exercice 92. Le tableau suivant donne le résultat du dosage du glucose sanguin, en grammes par litre, effectué sur un lot de 9 lapins :

1.17	1.16	1.15	1.18	1.19	1.20	1.16	1.20	1.21

On admet que chaque mesure est une variable normale d'espérance μ. Entre quelles limites peut-on situer μ au niveau de confiance 95% ?

Exercice 93. Une boisson de consommation courante est vendue en bouteilles d'un litre et contient une certaine quantité X d'un produit qui peut devenir toxique s'il est présent en grande quantité. On admet que X suit une loi normale de moyenne μ et que la boisson est conforme à la réglementation si la valeur de μ ne dépasse pas 50 mg/litre. Une association de consommateurs effectue un dosage de X sur un lot de 9 bouteilles et constate les concentrations suivantes en mg/litre :

60.5	58.5	57.8	56.4	54.7	53.5	49.3	48.2	48.0

Diriez-vous, au niveau d'erreur 5%, que le fabricant n'a pas respecté la réglementation ? Quelle erreur commettez-vous en affirmant que la réglementation n'est pas respectée ?

Exercice 94. On se propose de tester l'effet sur la pression artérielle d'un certain stimulus. On mesure à cet effet la pression artérielle de 6 sujets avant et après stimulus. On note par $(x_k)_{k=1}^6$ la pression avant le stimulus et par $(y_k)_{k=1}^6$ la pression après. On obtient les résultats suivants,

sujet	**1**	**2**	**3**	**4**	**5**	**6**
$\boldsymbol{x}$	12.4	11.8	12.8	13.5	13	12.7
$\boldsymbol{y}$	13.1	12.7	12.5	13.7	13.2	13.5

On suppose que les mesures d'un sujet à l'autre sont mutuellement indépendantes. Au niveau d'erreur 5 %, on demande si le stimulus agit sur la pression artérielle.

Exercice 95. Un relevé des hauteurs de pins, en mètres, dans deux forêts distinctes donne :

forêt 1	17.6	21.7	21.7	22.3
	23.1	24.6	24.7	26.3

forêt 2	16.5	16.5	18.6	19.4
	20.1	21.7	21.8	22.5
	22.9	23.3	23.6	24.1

Pensez-vous que les hauteurs moyennes diffèrent significativement d'une forêt à l'autre ?

Exercice 96. On traite 30 parcelles de terrain identiques avec deux types d'engrais différents. On mesure dans chacun des cas la production moyenne $\bar{x}$ et l'écart type s_{n-1} de la récolte. Dans les 15 premières parcelles, pour un engrais de type A, on trouve $\bar{x}_A = 3.6$ et $s_{n-1}(A) = 0.25$ quintal. Dans les 15 autres parcelles, pour un engrais de type B, on trouve $\bar{x}_B = 4.1$ et $s_{n-1}(B) = 0.27$ quintal. On supposera que les rendements sont indépendants d'une parcelle à l'autre et qu'ils suivent des lois normales $\mathcal{N}(\mu, \sigma^2)$.

1. En supposant que les écarts types soient identiques dans les deux cas, on montrera que les rendements ne diffèrent pas significativement d'un engrais à l'autre à 5 % près.
2. Déterminer la p-valeur du test de comparaison des moyennes.

Tests du chi-deux d'ajustement et d'indépendance

Exercice 97. On cherche à savoir si les naissances d'une certaine maternité se répartissent de manière uniforme tout au long de l'année. On dispose des données suivantes su 88 naissances :

Avril/Juin	**Juil/Août**	**Sept/Oct**	**Nov/Mars**
27	20	8	33

Que peut-on en conclure ? Si on regarde maintenant la répartition des naissances nationales tout au long de l'année, on constate :

Avril/Juin	**Juil/Août**	**Sept/Oct**	**Nov/Mars**
27 385	19 978	8 106	33 804

Que peut-on en conclure ? Déterminer aussi un intervalle de confiance à 95% de la proportion des naissances sur Avril/Juin.

Exercice 98. Lors d'une course de tiercé, on aligne 8 chevaux aux rangées numérotées de 1 à 8. La rangée numérotée 8 est la plus proche du centre du state. On se demande si un cheval a autant de chance de gagner indépendemment de sa rangée. Sur 110 courses de tiercé de chevaux de force équivalente, on rapporte dans le tableau suivant le nombre de fois qu'une rangée détient le gagnat.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
14	15	10	17	14	12	15	13

Les chances d'un cheval sont-elles les mêmes d'une place à l'autre ?

Exercice 99. On répertorie le nombre d'accidents de travail dans une certaine entreprise en fonction de l'heure de la journée. On trouve :

8 – 10	**10 – 12**	**13 – 15**	**15 – 17**
31	30	41	58

Peut-on affirmer au seuil d'erreur 10% que les accidents se répartissent uniformément au cours de la journée ? Quel est la p-valeur de ce test ?

Exercice 100. Un éditeur de presse cherche à établir un lien entre les ventes de trois quotidients A,B,C et le niveau social des acheteurs. Une enquête sur 300 lecteurs montre comment les niveaux sociaux professionnels se répartissent selon chaque quotidient. On obtient le tableau suivant.

	A	**B**	**C**
salariés	31	11	12
fonctionnaires	49	59	51
cadres	18	26	31
cadres supérieurs	2	4	6

Pensez-vous que le quotidient choisi dépend du niveau social du lecteur ? Jusqu'à quel seuil d'erreur, peut-on encore rejeter H_0 ?

Exercice 101. On désire savoir s'il existe, dans une population d'individus atteints du cancer de la peau, un lien entre l'âge de l'individu et ses chances de guérison. On mène une enquête sur trois classes d'âge et on obtient les résultats suivants :

âge\individu	**guéri**	**non guéri**
50 – 60 ans	1409	507
60 – 70 ans	763	248
70 – 80 ans	571	192

Pensez-vous que l'âge de l'individu est un facteur de guérison ?

Exercice 102. On répertorie dans le tableau suivant 300 accidents de voitures d'une année donnée suivant l'âge du conducteur et le nombre de ses contraventions enregistrées pendant les dix dernières années.

	contraventions		
	0	**1/2**	**2/+**
âge $\leqslant$ 21	8	23	14
22 $\leqslant$ âge $\leqslant$ 26	21	42	12
27 $\leqslant$ âge	71	90	19

Existe-t-il une corrélation entre l'âge du conducteur et le nombre de ses contraventions pendant les dix dernières années ?

Complément : exercices corrigés

Exercice 103. Une commune possède quatre pharmacies. La pharmatie A est la plus grande et la mieux située géographiquement. Pour cette raison, on pense que la moitié de la population se fournit en médicaments dans la pharmacie A et que l'autre moité se répartit équitablement entre les pharmacies B, C et D (la clientèle de cette commune étant particulièrement fidèle...). Afin de le vérifier, on sonde 200 habitants auxquels on demande où ils se procurent leurs médicaments. Voici les résultats obtenus :

Pharmacie *A*	**Pharmacie *B***	**Pharmacie *C***	**Pharmacie *D***
94	28	33	45

(1) Proposer un test. Préciser l'hypothèse nulle H_0 à tester.

(2) Faire fonctionner le test avec un risque d'erreur de 5% et conclure.

Exercice 104. Durant le premier semestre de l'année universitaire 2002-2003 un enseignant a effectué l'appel à chaque séance dans chacun des groupes de TD dont il avait la charge. Après les résultats des examens de février, l'enseignant a constitué le tableau ci-dessous ; en ligne, le nombre d'étudiants présents (chaque fois, presque toujours, quelque fois) ; en colonne, la fourchette des notes obtenues à l'examen :

	à chaque fois	**presque toujours**	**quelque fois**
entre 0 et 5	4	4	20
entre 6 et 10	23	15	12
entre 11 et 15	32	18	6
entre 16 et 20	15	7	4

Déterminer au risque $\alpha = 0.01$ si la présence en TD a eu une influence sur les résultats à l'examen.

Appendices

A Introduction au logiciel R

R est un logiciel de calcul scientifique orienté vers l'analyse des données en statistique. R est un logiciel du domain public comparable au logiciel professionnel S-Plus utilisé courament dans l'industrie et les laboratoires de recherche.

Il est fortement recommandé que chaque étudiant expérimente le logiciel R chez lui sur son ordinateur personnel. Le téchargement du logiciel doit se faire en mode administrateur. Le site officiel du Projet R est à l'adresse

http://www.r-project.org/

Il existe d'autres interfaces plus conviviales que celle fournie en standard par R. Il est conseillé de télécherger RStudio disponible sur http://www.rstudio.com/

A.1 Environnement de travail

Cette partie de l'appendice explique comment démarrer avec le logiciel R. La fenêtre principale est la console. Elle permet d'exécuter toutes les commandes de R, en particulier, les commandes de type unix de gestion de fichiers. Il est cependant fortement déconseillé de s'en servir pour l'exécution de longs programmes ou scripts. Nous verrons comment utiliser un script plus tard. En attendant, il est important de se familiariser avec les commandes de base du logiciel et comment obtenir de l'aide.

R utilisé en mode unix

La fenêtre console peut être utilisée en mode terminal de commandes. Par exemple

– Vérifiez bien le nom du répertoire courant avec

```
getwd()
```

– Eventuellement changer de répertoire

```
setwd("autre_repertoire")
setwd("../autre_repertoire")
```

"..." permet de remonter d'un cran l'arborescence des répertoires.

– Listez aussi tous les fichiers

```
list.files()
```

Il est fondamental de savoir comment obtenir rapidement des informations précises. L'aide en ligne de R est très riche et il ne faut hésiter à la consulter régulièrement.

– Depuis le menu, executez

→ Aide
→ Aide HTML
→ Search Engine & Keywords

puis écrivez dans le champ Search le nom d'une commande ou d'une fonction, par exemple

→ cos
→ base::Trig

Remarquez que cette méthode donne aussi des informations parallèles non pertinentes.

– Depuis le terminal, tapez la commande

```
?nom_fonction
```

```
?cos
```

Comparez avec la méthode précédente.

R dispose d'un nombre important de commandes ou fonctions qui sont chacune rangée dans des « packages ». La liste des packages s'obtient en tapant sur le terminal

```
library()
```

On obtient par exemple les packages base, graphics, tools, utils ... Pour des informations plus précises sur un packages, taper

```
library(help="base")
```

Remarquez que l'aide en ligne sur certaines fonctions nécessite une écriture différente. Par exemple, ?":" ou ?colon décrit la génération automatique de suites de pas ± 1. Une autre manière équivalente d'obtenir de l'aide est de taper la commande sous la forme

```
help("Extract") ; help("[[")
```

R utilisé en mode calculette

R peut être utiliser comme une simple calculette. Les opérations arithmétiques de base sont

?Arithmetic	
x + y	addition
x - y	soustraction
x * y	multiplication
x / y	division
x ^ y	puissance x^y
x %% y	reste $r = x \mod y$
x %/% y	dividende $(x - r)/y$

Les opérations de comparaison sont

?Comparison	
`x < y`	inférieur
`x > y`	supérieur
`x <= y`	inf ou égal
`x >= y`	sup. ou égal
`x == y`	égal
`x != y`	différent

Par exemple, les résultats de `1<2` et de `1>2` sont `TRUE` et `FALSE`. Les opérations logiques sont

?Logic	
`! x`	`NOT` logique
`x & y`	`AND` logique
`x | y`	`OR` logique
`xor(x,y)`	`OR` exclusif

Comme dans toute calculette, on retrouve l'ensemble des fonctions usuelles mathématiques

?S4groupGeneric			
`abs`	`sign`	`sqrt`	`ceiling`
`floor`	`trunc`	`round`	`signif`
`log`	`log10`	`log2`	`log1p`

Par exemple

```
ceiling(1.9)  # 2
floor(-1.9)   # -2
log10(1.0e3)  # 3
```

(Ne pas écrire `#` ni le résultat qui suit, car en fait `#` est le symbole sur `R` d'écrire des commentaires à la suite non exécutés). Remarquez comment un nombre en notation scientifique s'écrit :

```
5.1e-1 # égal à 0.51
-5.1e1 # égal à -51
```

Par défaut, les nombres sont affichés avec 7 chiffres décimaux. On peut modifier cet affichage par

```
getOption("digits") # 7
options(digits=20)
pi # 3.1415926535897993116
```

On dispose aussi des fonctions trigonométriques et hyperboliques

?Trig			
`acos`	`acosh`	`asin`	`asinh`
`atan`	`atanh`	`exp`	`expm1`
`cos`	`cosh`	`sin`	`sinh`
`tan`	`tanh`		

Par exemple

```
cos(pi/3)  # 0.5
atan(1)    # π/4 = 0.7853982
```

Remarquez que certaines constantes sont prédéfinies dans `R`, en particulier le nombre complexe $i = \sqrt{-1}$ (précédé d'un chiffre)

```
pi                  # 3.14
0.5+0.8660254i      # exp(iπ/3)
exp(1i*pi/3)        # exp(iπ/3)
1/2+(sqrt(3)/2)*1i  # 1/2 + √3/2 i
```

D'autres constantes non numériques existent aussi

```
LETTERS     # "A","B",...
letters     # "a","b",...
month.name  # "January",...
```

Les constantes numériques peuvent se présenter sous trois types différents : le type `double` (un réel), le type `complex` (un complexe) et le type `integer` (un entier). Par exemple

```
typeof(1)   # "double"
typeof(1i)  # "complex"
typeof(1L)  # "integer"
```

La manipulation des termes constituant un nombre complexe se fait par les fonctions

?complex				
`Arg`	`Conj`	`Im`	`Mod`	`Re`

Par exemple $\frac{\sqrt{2}}{2}(1+i) = \exp(i\frac{\pi}{4})$

```
Arg(sqrt(2)/2*(1+1i)
        )*360/(2*pi)    # 45
Mod(sqrt(2)/2*(1+1i))   # 1
```

On remarquera qu'un saut de ligne peut être introduit dans une instruction incomplète.

`R` dispose d'une riche collection de fonctions spécialisées. En voici quelques une

$$\Gamma(x) = \int_0^x t^{x-1}\exp(-t)\,dt, \quad x > 0,$$

$$B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x,y),$$

$$\binom{n}{k} = n(n-1)\cdots(n-k+1)/k!,$$

?Special	
`beta`	`lbeta`
`gamma`	`lgamma`
`digamma`	`trigamma`
`choose`	`lchoose`
`factorial`	`lfactorial`

Par exemple

```
choose(4,2)  # (4 2) = 6
gamma(4)     # Γ(4) = 3! = 6
```

`R` comme compilateur :

`R` n'est pas bien sûr uniquement une simple calculette. C'est un vrai langage de programmation au même titre que `C`, `Fortran` ou `java`, possédant en plus une bibliothèques de plusieurs milliers de fonctions (ou sous programmes) prêtes à l'emploi. Il est par exemple plus pratique d'écrire une seule liste d'instructions sur un fichier séparé appelé *script* et d'exécuter ce fichier en une seule fois, que de taper et d'exécuter chaque ligne de code l'une après l'autre directement sur le terminal.

Un fichier script permet aussi de sauvegarder l'ensemble du travail, de visualiser plus facilement les erreurs ou d'écrire de longs programmes. A partir de maintenant, il est donc recommander d'ouvrir un fichier script, de taper les lignes de code proposés, et de les excuter en revenant sur le terminal. La procédure exacte est la suivante

– Ouvrez un fichier vierge avec l'éditeur propre à R (tout autre éditeur *texte* conviendrait). Dans le menu allez sur

→ Fichier

→ Nouveau script

– Ecrivez dans le fichier script quelques lignes de code, par exemple un commentaire, puis à la suite deux lignes de code

```
# mon premier script
x <- 1+2+3+4
cat("x = ",x,"\n")
# ou plus simplement
print(x)
```

Remarquer bien, à la premieère ligne, qu'un commentaire commence toujours par # et que le texte qui suit n'est pas exécuté. La deuxième ligne est une affectation : x est une variable, <- est le symbole d'affectation. La troisième ligne permet d'afficher la valeur de x : cat pour *concatenate* ; "x = " et "\n" sont deux chaînes de caractères (la deuxième permet un retour à la ligne dans l'affichage). La fonction cat() permet un affichage précis tandis que la fonction print() est plus simple d'emploi.

– Enregistrez le fichier sous un nom, par exemple *mon_script.R* en n'oubliant pas l'extension .R. Retournez sur le terminal de commande en ligne et vérifiez que fichier se trouve bien dans votre répertoire courant avec list.files().

– Exécutez maintenant ce script en tapant sur le terminal de commande

```
source("mon_script.R")
```

Constater que l'affichage donne x = 10.

– Si l'utilisation de cat peut sembler fastidieuse pour des petits scripts, le simple fait d'écrire le nom d'une variable permet d'afficher son contenu. Par exemple le script

```
# mon premier script
x <- 1+2+3+4 ; x
```

permet sur une même ligne d'écrire deux instructions x <- 1+2+3+4 et x séparée par un point virgule. Il faut alors exécuter le fichier différemment

```
source("mon_script.R", echo=TRUE)
```

On retiendra enfin que, sur le terminal de commande, la touche ↑ rappelle la dernière commande et que → permet de completer automatiquement les noms de chemins et de fichiers.

A.2 Structures de données

R possède, comme tout langage de programmation puissant, des objets ou variables, des structures de données, des structures de contrôle et des fonctions à définir par l'utilisateur ou bien prêtes à l'emploi dans des librairies. R possède entre autre une collection très riche de fonctions graphiques et statistiques.

Les objets de base

– Le nom des objets ou variables doit suivre quelques règles. Il est formé des lettres A B ... Z, a b ... z, des chiffres 1 2 ... 9 et des deux caractères . et _ Mais il doit commencer par une lettre ou . et s'il commence par . il ne doit pas être suivi d'un chiffre. Par exemple

```
L1.bio   # est valable
_groupe  # n'est pas valable
._S_V_T  # est valable
.3ECTS   # n'est pas valable
```

– certains noms de variable sont réservés

?Reserved		
if	else	repeat
while	function	for
in	next	break
TRUE	FALSE	NULL
Inf	NaN	NA

comme les noms NA_integer, NA_real, NA_complex et NA_character ainsi que les noms ... ou ..1 ou ..2 qu'on retrouvent comme arguments dans les fonctions. La liste des objets est obtenue par objects().

– Chaque objet possède un type obtenu par typeof(*nom_objet*). Les principaux types sont

?typeof		
logical	integer	double
complex	character	raw

Par exemple

```
typeof(TRUE)  # "logical"
```

On notera que double est synonyme de numeric. Une chaîne de caractères s'écrit entre deux guillemets

```
mois_1 <- "janvier"
adverbe <- "aujourd\'hui"
```

Une chaîne caractères peut contenir des caractères spéciaux

?Quotes	
\n	*new line*
\r	*carriage return*
\t	tabulation
\\	*backslash*
\'	*ASCII apostrophe*
\"	*ASCII quotation mark*

– Tous les objets de R peuvent posséder zéro ou plusieurs attributs. La liste des attributs est accessible avec `attributes()` ; un attribut en particulier est accessible avec `attr()`. Les attributs possibles sont

```
"names", "dim", "dimnames", "class"
```

Les vecteurs

Les vecteurs sont des tableaux unidimensionnel d'objets de même type. Cette structure est dite *atomique*. C'est la structure de données la plus simple de R. Pour construire un vecteur, R utilise la construction *combine* `c` :

```
impair <- c(1,3,5,7,9)
pair <- 0:8  # 0,2,4,6,8
voyelles <- c("a","e","i","o","u","y")
```

– Il est possible de concaténer des vecteurs (de même type) par

```
chiffres <- c( c(0,1,2,3,4),
               c(5,6,7,8,9) )
```

– La génération de suite de nombres se fait aussi bien par l'opérateur « : »

```
s <- 0:5    # 0,1,2,3,4,5
s <- 0:(-5) # 0,-1,-2,-3,-4,-5
```

que par des constructions plus précises

?seq ?rep ?":"	
`seq()`	*sequence generation*
`rep()`	*replicate elements*

par exemple

```
x <- seq(from=0, to=5)
# x = 0,1,2,3,4,5
y <- seq(from=0, to=1, by=0.1)
# y = 0.0,0.1,...,0.9,1.0
z <- seq(from=0, to=-1,
         length.out=11)
# z = 0.0,-0.1,...,-1.0
```

ou bien

```
w <- c("a","b")
p <- rep(w, each=3)
# p = "a","a","a","b","b","b"
q <- rep(w, times=2)
# q = "a","b","a","b"
r <- rep(w, times=c(2,3) )
# r = "a","a","b","b","b"
```

(dans la dernière construction, la valeur de `times` doit être un vecteur de même longueur que celle de `w`). On peut aussi construire des vecteur logiques avec les opérateur de comparaison `< > <= >= == !=` et leurs combinaisons `! | xor` (voir les tableaux `?Comparison` et `?Logic` précédents). Par exemple

```
g <- c("A","O","A","AB")
s <- (g=="A")
# s = TRUE,FALSE,TRUE FALSE
h <- (-5):5
t <- (h>-3) & (h<=4)
# t = F,F,F,T,T,T,T,T,T,T,F
```

– L'indexation des vecteurs se fait à partir de 1 et utilise l'opérateur `[...]` où `...` peut être un vecteur d'entiers (tous positifs ou tous négatifs) ou un vecteur logique. Par exemple

```
s <- c("a","b","c")
# s = "a","b","c"
s[1]  # "a"
s[-1] # "b","c"
s[1] <- "z"
# s = "z","b","c"
s[-1] <- c("y","z")
# s = "a","y","z"
```

On peut aussi agir simultanément sur un ensemble d'indices. Par exemple

```
s <- 0:5     # s = 0,1,2,3,4,5
s[c(1,5)]    # 0,4
s[6:1]       # 5,4,3,2,1,0
s[(s%%2)==0] # 0,2,4
s[(s%%2)==1] <- c(0,2,4)
# s = 0,0,2,2,4,4
```

(dans les deux derniers exemples, le reste modulo 2, `s%%2`, est effectué sur chaque composante de `s` puis le test logique `(s%%2)==0` retourne un vecteur de même longueur que `s` permettant d'indexer les composantes paires de `s`).

– Pour préciser à l'avance la taille d'un vecteur d'un type donné, on utilise

```
logical(length=nn)
integerl(length=nn)
double(length=nn)
complex(length.out=nn, ...)
```

– Une des grandes forces des logiciels de calculs scientifiques est qu'ils peuvent réaliser des opérations arithmétiques terme à terme et en parallèle sur toutes les composantes d'un vecteur. Par exemple, les vecteurs numériques peuvent s'ajouter ou se multiplier terme à terme. Plus généralement, dans toute formule arithmétique, `cos(x)-y/z...`, les variables `x,y,z` peuvent être des vecteurs de longueur quelconque. Par exemple

```
(0:4)-1  # -1,0,1,2,3
c(2,4,6)/c(1,2,3)  # 2,2,2
(1:3)^2  # 1,4,9
```

```
cos(c(0,pi/3,pi/2))
# 1.0e+00 5.0e-01 6.1e-17
```

(le calcul exact donnerait $\cos(\pi/2) = 0$).

– Plusieurs fonctions agissent naturellement sur des vecteurs numériques, par exemple

?max ?min ...	
`max()`	*maximum*
`min()`	*minimum*
`prod()`	*produit*
`sum()`	*somme*

Le premier indice donnant le minimum ou le maximum est obtenu à l'aide de `which.min` et `which.max`. La fonction `which(x)` indique les indices des composantes vraies de `x`.

– Les fonctions suivantes agissent de manière paralèlle (terme à terme) ou cumulative

?pmax ?pmin ...	
`pmax()`	*maximum parallèle*
`pmin()`	*minimum parallèle*
`cummax()`	*maximum cumulatif*
`cummin()`	*minimum cumulatif*
`cumprod()`	*produit cumulatif*
`cumsum()`	*somme cumulative*

Par exemple

```
x <- pmax(1:6,6:1)
# x = 6,5,4,4,5,6
```

– Les fonctions `any(x)`, respectivement `all(x)`, disent si les composantes du vecteur logique `x` sont vraies pour l'une d'entre elles, respectivement pour chacune d'entre elles.

?any ?all	
`any()`	*au moins une composante*
`all()`	*toutes les composantes*

Par exemple

```
x <- seq(from=-5,to=5,by=2)
# x = -5,-3,-1,1,3,5
any(x==0)   # FALSE
all(x%%2==1)# TRUE
```

– Les vecteurs, comme tous les autres objets, possèdent plusieurs attributs. Le type et la longueur sont des *attributs intrinsèques*

?mode ?length	
`mode`	*comme* `typeof`
`length`	*taille du vecteur*

(sauf que le mode de `integer` et `double` est `numeric`). Les composantes d'un vecteur peuvent être nommées

```
vec <- 1:26
names(vec) <- LETTERS
vec["Z"] # le numéro 26
```

– `R` dispose d'une grande variété de focntions opérant sur les vecteurs :

`rev()`	*renverse l'ordre*
`sort()`	*réordonne*
`rank()`	*rang du réarrangement*
`match()`	*occurence d'un motif*
`append()`	*insère un vecteur*

Par exemple

```
vec <- append(c(1,2,3),c(-1,-1),2)
# vec = 1 2 -1 -1 3
```

Les facteurs

Les facteurs sont des vecteurs particuliers. Ils sont dits catégoriels car ils permettent de regrouper des données par catégorie. Deux facteurs permettent par exemple de paramétrer une table de contingence.

```
pop    <- c(100,200,100,300,200,200)
age    <- c(10,10,11,10,13,11)
sexe   <- c(0,1,0,1,0,1)
f.age  <- factor(age)
f.sexe <- factor(sexe)
liste  <- list(f.age,f.sexe)
tapply(pop,liste,sum)
```

La fonction `tapply()` permet d'appliquer la fonction `sum` aux individus dans `pop` regroupés par âge et par sexe. La table de contingence est alors :

```
     0   1
10 100 500
11 100 200
13 200  NA
```

(La fonction `table` est expliquée plus loin).

Les matrices et les tableaux

Les tableaux sont des vecteurs multi-indices. Les matrices sont des tableaux uniquement à 2 indices. Les deux fonctions `array(x,dim)` et `matrix(x,dim)` créent à partir du vecteur `x`, un tableau de dimensions extraites de `dim` en remplissant d'abord la première colonne. Le vecteur `x` est recyclé s'il n'est pas assez long. Pour la fonction `array()`, le vecteur `dim` peut contenir plus que 2 dimensions.

?array ?matrix	
`rarray`	2 ou plus d'indices
`matrix`	2 indices uniquement

L'exemple suivant donne un exemple de matrice de dimension `(2,3)` qui a été remplie à partir du vecteur `lettres` colonne par colonne en le recyclant autant de fois qu'il le faut.

```
lettres <- c("a","b","c")
taille  <- c(2,3)
tab <- array(lettres,taille)
```

Le résultat de `tab` est

```
     [,1] [,2] [,3]
[1,] "a"  "c"  "b"
[2,] "b"  "a"  "c"
```

– L'indexation se fait comme pour les vecteurs avec des crochets `[...]`

```
tab[2,3] # "c"
tab[2,]  # "b","a","c"
tab[,3]  # "b","c"
tab[c(1,2),c(2,3)]
```

qui donne comme résultat

```
     [,1] [,2]
[1,] "c"  "b"
[2,] "a"  "c"
```

On retiendra la construction très pratique pour obtenir une ligne : `tab[2,]` (la ligne 2), ou une colonne : `tab[,3]` (la colonne 3). La fonction `matrix()` construit uniquement des tableaux 2-dimensionnels. L'exemple suivant remplit un tableau par ligne,

```
matrix(1:12,nrow=4,ncol=3,byrow=TRUE)
```

qui donne comme résultat

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
[4,]   10   11   12
```

```
"cbind"
colSums
rowSums
colMeans
rowMeans
```

Les listes

Les listes permettent de collecter dans un même objet des données qui ne sont pas du même type. Chaque donnée est une structure comme un vecteur, un facteur, un tableau, mais aussi une liste. Pour accéder à chaque partie de la liste on utilise l'opérateur `$`

```
fichier <- list(
taille=c(1.7,1.7,1.8,1.6,1.8,1.7,1.5),
groupe=c("O","A","A","AB","A","O" ))
fichier$taille # 1.7,1.7,1.8, ...
fichier[[2]]   # "O" "A", "A"...
# pour modifier une valeur
fichier$groupe[4] <- "O"
length(fichier$taille) # 7
length(fichier$groupe) # 6
```

A.3 Gestion des entrées/sorties

Les entrées/sorties se divisent en deux catégories : celles de haut niverau et celles de bas niveau.

`read.table()`

est la manière la plus simple de créer un *data.frame* à partir d'un fichier. Les autres entrées/sorties sont

```
"file" "close" "writeLines"
"readLines" "read.table"
"scan"
```

A.4 Outils de statistique

```
"table" "mean" "sample"
"median" "var" "cov" "cor"
```

– `R` supporte un grand nombre de distributions statistiques. Chacune de ces distributions se présente sous quatre formes. Si *distrib* est l'une de ces distributions, on a

`d`*distrib*`()` : densité $f(t)$
`p`*distrib*`()` : fonction de répartition
`q`*distrib*`()` : quantile q_α d'ordre α
`r`*distrib*`()` : n nombres aléatoires

Si $f(t)$ désigne la densité d'une variable aléatoire continue X, la fonction de répartition est $\mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(t)dt$, le quantile q_α est $\alpha = \mathbb{P}(X \leqslant q_\alpha)$ et `r`*distrib* génère n nombres aléatoires de loi X. `R` fournit un grand nombre de telles distributions ; par exemple

`?Distributions` distributions discrètes	
`dbinom`	binomilale
`dgeom`	géométrique
`dpois`	Poisson

`?Distributions` distributions continues	
`dbeta`	beta
`dcauchy`	Cauchy
`dexp`	exponentielle
`df`	Fisher ou f
`dgamma`	gamma
`dnorm`	normale
`dt`	Student ou t
`dunif`	uniforme

Par exemple $p = \mathbb{P}(Z \leqslant 1.5)$ donne

```
p <- pnorm(1.5, mean=0, sd=1)
# p = 0.9382198
```

$\mathbb{P}(|Z| \geqslant q) = 0.13$ donne

```
q <- qnorm(1-0.13/2, mean=0, sd=1)
# q = 1.514102
```

B Modélisation sur logiciels

Cet annexe propose 4 travaux pratiques sur le logiciel `R`. Les TP font parties de la note de contrôle continu. Les étudiants peuvent choisir de travailler en binôme ou en trinôme. Chaque groupe d'étudiants rend un unique fichier script par courrier électronique aux enseignants correspondants.

`R` est un logiciel de calcul scientifique orienté vers l'analyse des données en statistique. `R` est un logiciel du domain public comparable au logiciel professionnel `S-Plus` utilisé courament dans l'industrie et les laboratoires de recherche.

B.1 Introduction au logiciel R

Ce TP ne sera pas à rendre et ne donnera pas lieu à une note.

Environnement de travail. Avant toute utilisation de `R`, il est nécessaire de bien s'organiser.

– Lancez le logiciel `R`. Cela dépend du type de système d'exploitation, Unix, Windows, Mac, et de l'interface utilisée. Une première fenêtre s'ouvre appelée `Console`.

– Il est important de de choisir un répertoire de travail dans lequel les images et fichiers seront sauvegardés. Allez dans le menu, puis

`Fichier → Changer le répertoire courant ...`

– Vérifiez bien le nom du répertoire courant par `getwd()`. Listzez tous les fichiers, `list.files()`. Eventuellement pour changer de répertoire, on peut utiliser `setwd()` ; par exemple

`setwd("../`*nouveau_répertoire*`")`

– `R` peut être utilisé interactivement depuis la console, mais il est plus commode d'écrire un script séparément et de l'exécuter par la suite sur la console. C'est ce qu'on fera par la suite. Allez dans le menu :

```
Fichier -> Nouveau script
```

pour créer un fichier vierge.

– Ecrivez quelques lignes de commentaires : nom du tp, nom des étudiants en binôme ou trinôme

```
# TP1
# noms, prenoms : ...
```

Remarquer bien qu'un commentaire commence toujours par `#`; le reste de la ligne n'est alors pas exécutée. Habituez-vous à commenter votre code de programmation.

– Enregistrez le fichier sous le nom `TP1.R` en n'oubliant pas l'extension `.R`. Vérifiez qu'il se trouve dans votre répertoire courant : `list.files()`

– Testez une opération simple : d'abord sur la console en tapant par exemple

```
choose(4,2)
```

puis tapez un retour à la ligne. La commande calcule le coefficient binomial $\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4*3}{2*1} = 6$. Le résultat 6 s'affiche.

– Recommencez cette opération dans le script `TP1.R` : écrivez sur deux lignes séparées sans mettre les commentaires

```
x <- choose(4,2) # affectation d'une variable x
cat(x) # affichage de x
```

Exécutez maintenant ce script en revenant sur la console et en tapant la commande

```
source("TP1.R")
```

Constatez bien que le résultat 6 est affiché. On notera que, sur la console, la touche ↑ rappelle la dernière commande et que → permet de completer automatiquement les noms de chemins et de fichiers.

– Pour obtenir de l'aide en ligne sur une commande : écrivez sur la console `?`*`nom_commande`*. Par exemple taper

```
?choose
```

(Un autre méthode plus simple, pour un système Unix seulement, consiste, dans le fichier script, à mettre le pointeur sur le nom de la fonction puis à cliquer sur `F2`).

– Pour obtenir de l'aide en ligne en général, depuis le menu, exécuter

```
Aide -> Aide HTML
```

Puis cliquer sur `Search Engine & Keywords` et écrire `Choose` dans le champ `Search`. Comparer avec la méthode précédente.

Les bases de R

A partir de maintenant, vous taperez toutes les commandes dans le fichier script `TP1.R` *puis vous l'exécuterez en reveant sur la console et en tapant* `source("TP1.R")` *On utilisera les flêches ↑ pour rappeler la commande précédente, par exemple* `source("TP1.R")` *et → pour compléter automatiqument les noms de fichier.*

L'annexe A, que vous devez avoir lue, contient toutes les informations utiles pour comprendre le maniement du logiciel `R`*. Le seul objectif de ce TP est de vous familiariser avec son langage de programmation. Les TP suivants supposeront que les bases de* `R` *sont acquises.*

– L'affectation d'une variable se fait avec `<-`. Expérimentez

```
n1 <- pi
n2 <- exp(1)
n3 <- 1/4
n4 <- 3^3
n5 <- "Aa"
cat(n1, "\n", n2, "\n", n3, "\n", n4, "\n", n5, "\n")
```

Pour afficher les valeurs des variables sur la console, on utilise dans le script la commande `cat` (sur la console, il aurait suffi de taper le nom de la variable). Le caractère `\n` désigne un retour à la ligne. Les guillemets `"..."` désigne une chaîne de caractère. Le type de la variable s'appelle `mode`. Tapez sur la console

```
mode(pi)
mode("Aa")
```

– La structure de données la plus simple est la structure `vector`

```
m1 <- c(-1,1,-2,2) ; m3 <- 1/m1
m2 <- c("A", "a", "B", "b", "C", "c")
m4 <- c(m1, m3, m1*m3)
cat(m1, "\n", m2, "\n", m3, "\n", m4, "\n" )
```

A chaque fois, il faut bien comprendre le sens de l'opération. Que fait `m3` ? Que fait `m4` ? Remarquez aussi qu'on peut séparer deux instructions par `;`

– Les opération arithmétiques se font composante par composante sur chaque élément des vecteurs. Les deux opérateurs suivant `:` et `seq` créent des suites de nombres de pas constant (pour `x:y` le pas est égal a 1). Expérimenter

```
p1 <- ((-10):10)/10
p2 <- seq(from=-1, to=1, by=0.1)
p3 <- p1-p2
p4 <- p1/p2
p5 <- sum(p1)
cat(p1, "\n", p2, "\n", p3, "\n", p4, "\n", p5, "\n")
```

C'est le moment d'aller chercher l'aide en ligne : tapez sur la console

```
?sum
?":"
?seq
```

puis lisez les pages en entier, même si c'est ardu.

– On peut arranger une série de nombres sous forme de tableau multidimensionnel appelé `array`. Expérimentez sans écrire les commentaires qui sont présents uniquement pour vous aidez. Au lieu de `cat()`, le code utilise `print()` ; à vous de choisr ce que vous préférez.

```
q0 <- array(1:9, dim=c(3,3)) # un tableau 3x3
q1 <- rep(c(1,2),times=3) # répétition de c(1,2)
q2 <- array(q1, dim=c(2,3)) # un tableau 2x3
q3 <- array(q1, dim=c(2,3),
      dimnames=list(c("L1","L2"),c("C1","C2","C3")))
q4 <- q2[2,3] # indexation par des nombres
q5 <- q3["L1",] # indexation par des noms
print(q0) ; print(q1) ; print(q2)
print(q3) ; print(q4) ; print(q5)
# l'affectation de la colonne C3
q3[,"C3"] <- c(0,0)
```

```
print(q3)
```

Remarquez bien comment le vecteur `1:9` a été aligné dans le tableau `q0`. On constate aussi que l'indexation peut se faire aussi bien avec des nombres qu'avec des noms `q3["L2","C3"]` (à condition qu'on ait nommé les colonnes et lignes auparavant). Remarquer enfin qu'on peut extraire des lignes complètes `q2[1,]` ou des colonnes complètes `q2[,3]`.

– La fonction de base pour afficher un graphique est `plot`. On commence par un exemple simple de superposition de 3 densités de loi normale

```
x <- seq(from=-10, to=10, by=0.01)
y <- dnorm(x, mean=1, sd=1)
z <- dnorm(x, mean=2, sd=1.5)
w <- dnorm(x, mean=3, sd=2)
plot(x,y, type="l", col="green")
lines(x,z, type="l", lty=2, col="blue") # il s'agit de ℓ et non pas de 1 dans type
lines(x,w, type="l", lty=3, col="red")
legend(-10, 0.4, legend=c("mean=1, sd=1","mean=2, sd=1.5","mean=3, sd=2"),
  lty=c(1,2,3), col=c("green", "blue", "red"))
```

– Recherchez d'abord l'aide en ligne de

```
?dnorm # à taper sur la console
```

Expliquez ce que fait cette commande. Puis recherche l'aide en ligne de `?plot`, `?lines` et `?legend`. Si vous manquez d'idée pour une couleur, tapez sur la console `colors()`.

– Expérimentez maintenant l'exemple plus compliqué suivant sans écrire les commentaires. Il servira au TP prochain.

```
r1 <- seq(from=-4, to=4, by=0.01)
r2 <- dnorm(r1) # densité de la loi normale
r3 <- dt(r1,6) # densité de la loi de Student : ddl=6
r4 <- dt(r1,60) # densité de la loi de Student : ddl=60
r5 <- dbinom(0:10,10,0.5) # loi binomiale : n=10 et p=0.5
oldpar <- par(no.readonly = TRUE) # ancienne configuration
par(mfrow=c(1,3)) # 3 figures sur une même ligne
# dans le plot suivant type="l" ; ell pour ligne
plot(r1,r2, type="l", col="green", main="Loi Z normale",
    xlab="x", ylab="densité en x")
plot(r1,r3, type="l", main="Loi T de Student",
    xlab="x", ylab="densité en x")
par(new=TRUE) # superposition des figures
plot(r1,r4, type="l", col="blue",
    xlab="", ylab="", xaxt="n", yaxt="n")
par(new=FALSE) # suppression de la superposition
plot(0:10,r5, type="h", # h pour histogramme
    main="loi B binomiale", xlab="k", ylab="P(B=k)", lwd=3)
par(oldpar) # on récupère l'ancienne configuration
```

Il est important ici de lire la documentation concernant les graphiques :

```
?plot ?plot.default ?par ?hist ?barplot.
```

On devra trouver la figure 1.

B.2 Modélisation probabiliste

Les TP font partie de la note de contrôle continu. Chaque étudiant (ou groupe d'étudiant) rend un unique fichier script par courrier électronique aux enseignants correspondants.

Echantillon de loi discrète donnée
Partie du TP à ne pas rendre.

On considère une variable aléatoire X prenant r valeurs $(\xi_1, \ldots, \xi_r)$ de probabilités $p = (p_1, \ldots, p_r)$. On cherche à construire une fonction sous le même modèle que `rnorm`, qu'on utilisera par la suite sous la forme

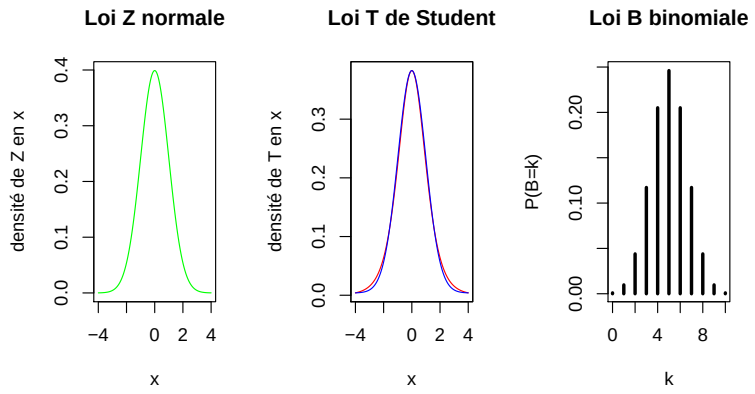


FIGURE 1 – Différentes distributions et leur loi

```
echantillon <- rdiscrete(n,p,xi)
```

et donnant en sortie un échantillon de taille `n`, de loi `p` et de modalités `xi`.

– Toute les fonctions suivent une syntaxe similaire à ce qui suit

```
rdiscrete <- function(n,p,xi)
{ # début de la fonction
# n = entier
# p = c(p_1,...,p_r) où p_i = nombre
# xi = c(xi_1,...,xi_r) où xi_i = caractère
...
...
return(echantillon) # résultat de la fonction
} # fin de la fonction
```

Il reste bien sûr à compléter la partie ... c'est-à-dire à évaluer `echantillon` en fonction des paramètres `n`, `p` et `xi`. L'idée est de découper l'intervalle $[0,1]$ selon les proportions p_i, puis de lancer au hasard n points uniformément sur cet intervalle. S'ils tombent dans la i-ième tranche, c'est la modalité ξ_i qui apparaît. La solution est donnée en intégralité dans le code suivant (ne pas écrire les commentaires !) Comprenez bien ce code car il pourra être utile plus tard.

```
# on crée un vecteur générique de taille n et de type xi
echantillon <- vector(mode=mode(xi), length=n)
# on génére un échantillon uniform e_unif[k] de taille n
e_unif <- runif(n)
# on calcule le nombre de modalités r
r <- length(p)
# on crée les points de subdivision de l'intervalle [0,1]
pp <- c(0,cumsum(p))
# on fait tourner une boucle
for(i in 1:r) {
# on crée un vecteur logique de taille n qui contient TRUE
# à l'indice k si e_unif[k] tombe dans la i-ième tranche
ii <- (pp[i] <= e_unif) & (e_unif < pp[i+1])
# on index un vecteur par un vecteur logique : aux indices
# TRUE echantillon prend la modalité xi[i]
echantillon[ii] <- xi[i]
# fin de la boucle
}
```

– Le groupe sanguin se répartie chez les Basques selon des proportions

$$p_O = 56\%,\ p_A = 40\%,\ p_B = 3\%,\ p_{AB} = 1\%$$

légèrement différentes de celles de la polulation française en général de

$$p_O = 43\%,\ p_A = 45\%,\ p_B = 9\%,\ p_{AB} = 3\%$$

– Créez d'abord deux échantillons de taille $n = 1000$ pour chaque groupe sanguin. On donne encore la solution

```
xi <- c("0", "A", "B", "AB")
p_basque <- c(0.56, 0.40, 0.03, 0.01)
p_nation <- c(0.43, 0.45, 0.09, 0.03)
n <- 1000
e_basque <- rdiscrete(n,p_basque,xi)
e_nation <- rdiscrete(n,p_nation,xi)
```

– On factorise chaque échantillon selon les modalité `"O", "A", "B", "AB"` (non rangées dans l'ordre lexicographique) puis on construit une table de contingence (à une variable ici)

```
f_basque <- factor(e_basque, levels=xi)
f_nation <- factor(e_nation, levels=xi)
t_basque <- table(f_basque)
t_nation <- table(f_nation)
```

– Afficher des informations pour mieux comprendre

```
print(t_basque)
print(t_nation)
```

– On réunit les deux lignes en un seul tableau et on affiche le diagramm en bâton

```
groupe_sanguin <- rbind(t_basque/n, t_nation/n)
barplot(groupe_sanguin, beside=TRUE, ylim=c(0,0.6),
    legend.text=c("groupe sanguin basque",
    "groupe sanguin national"))
```

On doit trouver la figure 2

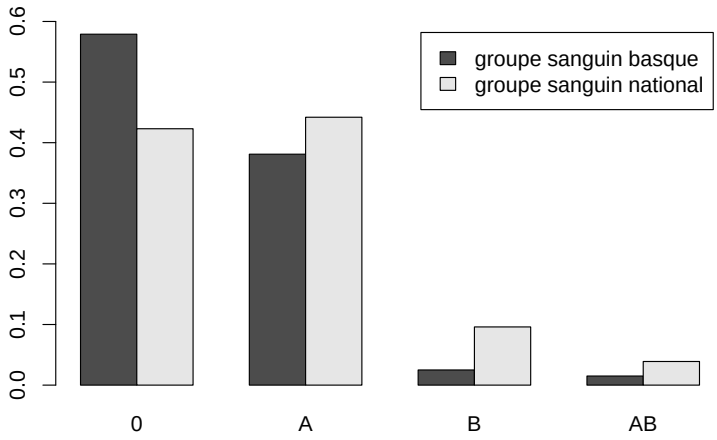


FIGURE 2 – Diagramme en bâton des deux groupes sanguins

Le théorème central limite

Partie du TP à rendre. N'oubliez pas de commencer votre script par `# TP2.R` *et les noms de chaque étudiant :* `# noms, prenoms` *formant le binôme ou le trinôme.*

On cherche à montrer graphiquement pourquoi la distribution d'une loi binomiale centrée réduite

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}, \quad Y_n \overset{loi}{\sim} \mathcal{B}(n,p), \quad \mathbb{E}[Y_n] = np, \ \mathrm{Var}(Y_n) = np(1-p)$$

converge en loi vers la loi normale $\mathcal{N}(0,1)$. On constate d'abord que Z_n prend les valeurs disctrètes

$$z_{n,k} = \frac{k - np}{\sqrt{np(1-p)}}, \quad \text{avec probabilité} \quad \mathbb{P}(Z_n = z_{n,k}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

On constate ensuite que le pas entre deux $z_{n,k}$ successifs est égal à

$$\Delta_n = 1/\sqrt{np(1-p)}.$$

On définit ainsi une densité $f_n(z)$ constante entre deux $z_{n,k}$ successifs par la formule

$$f_n(z)\Delta_n = \mathbb{P}(Z_n = z_{n,k}), \quad \forall\, z \in]z_{n,k-1}, z_{n,k}].$$

– On demande de représenter sur un même graphique les trois distributions $n = 10$, $n = 100$ et $n = 1\,000$ pour $p = 0.5$ (figure 3) et pour $p = 0.05$ (figure 4). On superposera aussi à chaque fois la distribution de la loi normale centrée réduite.

On remarque que la convergence de Z_n vers $\mathcal{N}(0, 1)$ est beaucoup plus lente lorqsque p est proche de 0 ou de 1 : pour un échantillon de taille $n = 100$ et pour $p = 0.05$, l'erreur commise entre les deux distributions est très importante.

– Indication : Faire simultanément les cas $p = 0.5$ et $p = 0.05$. On découpera l'espace graphiqe en 6 cases avec

```
par(mfrow=c(2,3))
```

Pour tracer l'histogramme de Z_n on tracera deux fois les graphes de $f_{n,k} = \mathbb{P}(Z_n = z_{n,k})/\Delta_n$ en fonction de $z_{n,k}$ selon le modèle suivant

```
# créer l'échantillon z_nk et la distribution f_nk
plot(z_nk, f_nk, type="h", xlim=c(-3,3), ylim=c(0,0.4) )
par(new=TRUE)
plot(z_nk, f_nk, type="s", xlim=c(-3,3), ylim=c(0,0.4) )
```

On superposera aussi la courbe de la densité de la loi normale (aller voir l'aide en ligne de `dnorm`)

```
# créer un échantillon z_norm de [-3,3] de pas 0.01
# calculer la densité f_norm de la loi normale en z_norm
par(new=TRUE)
plot(z_norm, f_norm, type="l", xlim=c(-3,3), ylim=c(0,0.4))
```

Ou pourra compléter les graphiques par de la couleur `col=...`, un titre `main=...`, des noms d'axe `xlab=...`, `ylab=...`

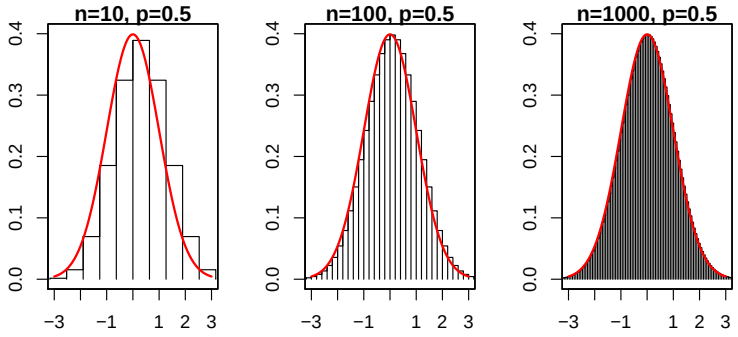


FIGURE 3 – Distribution de la loi binomiale symétrique

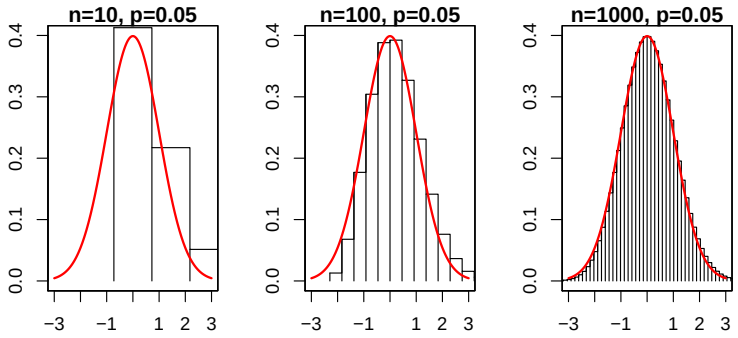


FIGURE 4 – Distribution de la loi binomiale excentrée

B.3 Statistique inférentielle : intervalles de confiance

Le TP en entier est à rendre. N'oubliez pas de commencer votre script par `# TP3.R` *et les noms de chaque étudiant :* `# noms, prenoms` *formant le binôme ou le trinôme.*

Dans toute la suite, on considère le poids (fictif) de 30 nourrissons dans une maternité A et de 50 autres nourrissons dans une maternité B :

```
poidsA <- c(rnorm(15,3050,450),rnorm(15,3400,250))
poidsB <- c(runif(20,2600,4500),runif(30,3150,3650))
```

Représentation graphique

– Affichez une figure contenant les deux histogrammes en indiquant bien en sur-titre les deux maternités. Utilisez `hist` et son aide en ligne. On devra trouver un figure semblable à la figure 5. (On laissera `R` choisir le nombre de colonnes).

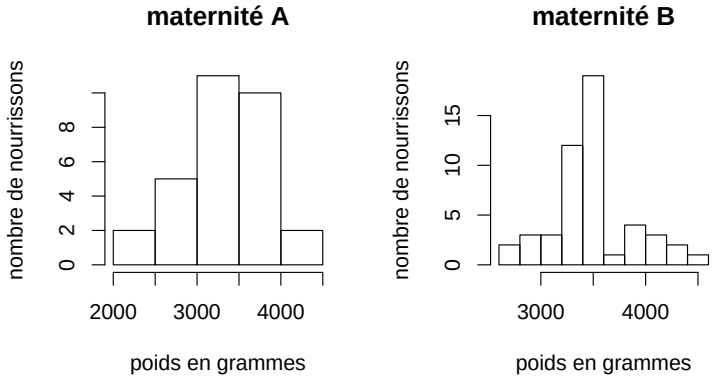


FIGURE 5 – Histogramme des naissances

– Afficher les quantiles de chaque distribution : on pourra essayer les deux méthodes

```
cat("quantiles de A : \n")
print(quantile(poidsA))
print(summary(poidsA))
```

– Remarque : pour arrêter le défilement des figures à l'affichage, on insérera en début de script la commande `devAskNewPage(ask = TRUE)`. Pour stopper l'exécution du script à tout moment, il peut être util d'utiliser `scan()`. La reprise de l'exécution du programme se fait par un retour à la ligne (sur PC) ou `OK` (sur Unix).

– Affichez les boxplots des 2 distributions sur un même graphique. On utilisera la commande `boxplot` et ses options `range` et `names`. Il est impératif d'aller voir l'aide en ligne de `boxplot`. Indiquer bien le nom de chaque boxplot A ou B ; indiquer aussi l'option choisie sur `range`. On devra trouver une figure semblale à la figure 6.

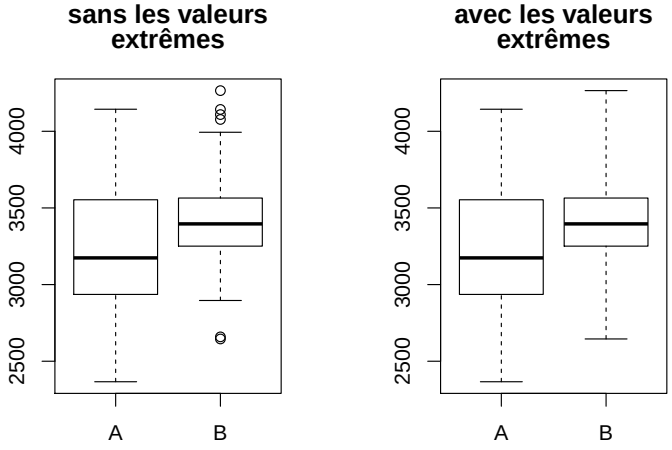


FIGURE 6 – Boxplot des naissances dans deux maternités

Intervalle de confiance

– En utilisant les formules donnant l'intervalle de confiance de la moyenne μ_A des poids de la maternité A, écrivez un code de programme donnant les extrémités $[\mu_{min}, \mu_{max}]$ au seuil 2%. On rappelle la formule

$$\mu_{min} = \bar{x} - t_\alpha \frac{s_{n-1}}{\sqrt{n}} \leqslant \mu_A \leqslant \bar{x} + t_\alpha \frac{s_{n-1}}{\sqrt{n}} = \mu_{max}.$$

où $t_\alpha = q_{1-\alpha/2}$ et q_β est le quantile d'orde β de la loi de Student à $n-1$ degrés de liberté. On utilisera pour cela les commandes `mean` (moyenne), `sd` (écart-type), `qt` (quantile de la loi de Student). On complétera les `...` du code suivant

```
muA <- mean(poidsA)
sigmaA <- sd(...)
alpha <- 0.02
nA <- length(...)
t_alpha <- qt(...,...)
mu_min <- ... # formule du cours
mu_max <- ... # formule du cours
cat("IC de muA : formule du cours\n")
cat("mu_min = ", mu_min, "\n")
cat("mu_max = ", mu_max, "\n")
```

`R` fait bien sûr tous ces calculs. Comparez les résultats précédents avec ceux du code suivant. Lisez d'abord l'aide de `t.test`. Le résultat est une structure de type `list` ; les variables d'une liste sont accessibles avec l'oprérateur `$`.

```
icA <- t.test(poidsA, conf.level=1-alpha, alternative="two.sided")
mu_min_bis <- icA$conf.int[1]
mu_max_bis <- icA$conf.int[2]
cat("IC de muA : t.test \n")
cat("mu_min_bis = ", mu_min_bis, "\n")
cat("mu_max_bis = ", mu_max_bis, "\n")
```

– Déterminez un intervalle de confiance de $\mu_B - \mu_A$: on utilisera à nouveau `t.test` avec l'option `var.equal=TRUE` (donnant les formules du cours) puis avec `var.equal=FALSE` (non développé dans le cours). Formulez par exemple les résultats sous la forme

```
icAB <- t.test(...)
mu_min_AB <- ... # utilise icAB
mu_max_AB <- ... # utilise icAB
cat("IC de muB-muA : var.equal=TRUE \n")
cat("mu_min_AB = ", mu_min_AB, "\n")
cat("mu_max_AB = ", mu_max_AB, "\n")
```

S'agit-il d'un intervalle de confiance dans le cas apparié ou dans le cas indépendant ? Il est indispensable d'aller regarder l'aide en ligne de `t.test` et de son option `paired`. Répondez par

`cat("interval dans le cas...\n") # apparié ou indépendant`.

– Pour les intervalles de confiance d'une proportion on utilise `binom.test`. Dans un échantillon de 197 pommes, on constate que 19 d'entre elles sont abîmées. Déterminez un intervalle de confiance de la proportion de pommes abîmées. On utilisera d'abord les formules du cours

$$p_{min} = \hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqslant p \leqslant \hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = p_{max}$$

```
n_tot <- 197
n_obs <- 19
alpha <- 0.05
z_alpha <- qnorm(...)
p_min <- ... # remplacer ... par les formule du cours
p_max <- ... # remplacer ... par les formule du cours
cat("IC de p : formule du cours \n")
cat("p_min = ", p_min, "\n")
cat("p_max = ", p_max, "\n")
```

Recommencez avec `binom.test` et comparer les résultats

```
icp <- binom.test(...)
p_min_binom <- ... # utilise icp
p_max_binom <- ... # utilise icp
cat("IC de p : binom.test \n")
cat("p_min_binom = ", p_min_binom, "\n")
```

```
cat("p_max_binom = ", p_max_binom, "\n")
```

Comparez ces derniers résultats avec ceux obtenus avec les formules du cours.

B.4 Statistique inférentielle : tests d'hypothèse

Le TP en entier est à rendre. N'oubliez pas de commencer votre script par `# TP4.R` *et les noms de chaque étudiant :* `# noms, prenoms` *formant le binôme ou le trinôme.*

On s'intéresse au poids d'un lot de 200 yaourts au caramel et de 200 yaourts au chaocolat. On cherche dans une première partie à vérifier si le poids des yaourts au caramel est conforme à l'affichage (poids supérieur à 100 g). Dans une deuxième partie, on compare dans un test d'hypothèse les poids moyens des deux types de yaourts. On cherche enfin dans une troisième partie à vérifier si la distribution des poids des yaourts au chocolat est gaussienne.

Récupération d'un fichier sur internet

– Récupérez un fichier de données sur internet par la commande suivante

```
adresse_fichier <-
      "http://www.math.u-bordeaux1.fr/~thieulle/Data/2012_001.txt"
poids <- read.table(adresse_fichier, header=TRUE, sep="", quote="", dec=",")
```

Bien comprendre l'aide en ligne de `?read.table` et ses paramètres `header` pour une première ligne de texte, `sep` pour la séparation des colonnes, `quote` pour les données non numériques, `dec` pour la convention des nombres décimaux. Le résultat `poids` est un `data.frame` ou une liste de vecteurs de même longueur.

Extraction de données du fichier

Le fichier donne le poids de petits pots de caramel ou de chocolat. Extraire d'abord quelques informations

```
cat("affichage des 10 premières lignes : \n")
print(poids[1:10,])
cat("noms des colonnes : \n")
print(names(poids))
cat("nombre de petits pots : \n")
print(length(poids[,1]))
cat("résumé statistique : \n")
print(summary(poids))
```

– Affichez sur une même figure les deux histogrammes (on rappelle qu'on accède aux objets d'une liste par `$`)

```
hist(poids$caramels, ...)
hist(poids$chocolats, ...)
```

On devra trouver la figure 7 et on n'oubliera pas d'annoter les graphiques. Remarquez comme les deux histogrammes sont différents.

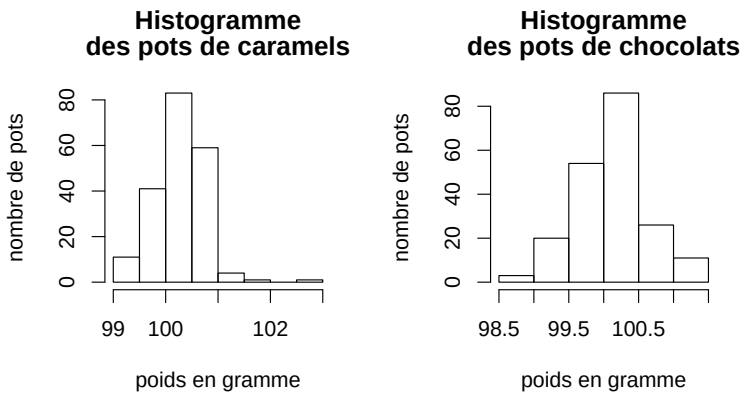


FIGURE 7 – Histogramme des poids de petits pots lactés

Test de conformité des pots de caramel

Le fabriquant produit par heure 200 pots et garantit que le poids dans chacun des cas est de 100 g au moins. Faites un test sur l'hypothèse alternative « le poids est supérieur à 100 g » d'abord sur tous les pots produits pendant une heure (tout le fichier). Utilisez `t.test()` en allant voir l'aide en ligne et formatez les réponses sous la forme

```
alpha <- 0.005
test_caramels <- t.test(poids$caramels, ...)
cat("test sur 200 pots de caramels : \n")
print(test_caramels)
```

Quelle est la p-valeur du test ?

– Il est possible que le contrôle de ces pots soit destructif (le fabriquant contrôle la composition en même temps). Le fabriquant ne peut pas se permettre alors de détruire la production d'une heure. Il choisit au hasard 20 pots parmis ces 200 et fait le test sur cet échantillon. Pour choisir au hasard, utilisez `sample()` et son aide en ligne.

```
poids_caramels <- sample(poids$caramels, ...)
alpha <- 0.005
test_caramels <- t.test(poids_caramels, ...)
cat("Test sur 20 pots de caramels : \n")
print(test_caramels)
```

Remarquez maintenant comme la p-valeur a augmenté. Expliquez pourquoi.

– Pour bien comprendre ce que fournit `t.test`, calculer la p-valeur de l'échantillon de 20 pots en utilisant la formule du cours

$$t_{crit} = \sqrt{n}\,\frac{\bar{x} - 100}{s_{n-1}}, \quad p_{valeur} = \mathbb{P}(\mathcal{T} > t_{crit}).$$

où $\mathcal{T}$ est une loi de Student de $ddl = n-1$. On écrira les calculs sous la forme suivante et on comparera avec le dernier `print(test_caramels)`.

```
t_crit <- ...
p_valeur <- ...
cat("p-valeur du cours : \n")
cat("t_crit : ", t_crit, "\n")
cat("p-valeurs : ", p_valeur, "\n")
```

Faites tourner cette dernière partie du programme (y compris le choix des 20 pots de caramels) plusieurs fois en remarquant que t_{crit} et p_{valeur} changent à chaque fois. Comment l'expliquez-vous ?

Test de comparaison des moyennes des poids

On réalise maintenant un test d'hypothèse sur la différence des moyennes des poids. On prend là aussi, d'abord les 200 pots, puis deux échantillons de 20 pots de caramel et chocolat pris au hasard. On choisit entre le cas apparié et le cas indépendant. Formulez par exemple les réponses sous la forme

```
poids_chocolats <- sample(poids$chocolats, ...)
alpha <- 0.005
test_poids <- t.test(poids$caramels, poids$chocolats, ...)
cat("Test de différence de moyenne sur 200 pots : \n")
print(test_poids)
test_poids <- t.test(poids_caramels, poids_chocolats, ...)
cat("Test de différence de moyenne sur 20 pots : \n")
print(test_poids)
```

On notera que la probabilité d'affirmer à tort que les moyennes des poids diffèrent dans les deux lots, est égale à 1.4 sur mille pour un échantillon de 200 pots, alors qu'elle peut être supérieure à plus de 70% pour un échantillon de 20 pots.

Comparaison à une distribution gaussienne

L'histogramme des poids des 200 pots de chocolats suggère que la répartition est gaussienne. On cherche à réaliser un test du chi-deux d'ajustement. On commence par créer un tableau de poids par classes. On prendra

$$]98.5, 99], \quad]99, 99.5], \quad \ldots \quad ,]101, 101.5].$$

On utilisera `seq` pour générer les classes, et `cut` pour récupérer les effectifs par classes. Le tableau des effectifs par classes est donné par la nouvelle structure `table`. La fonction `levels` récupère le nom des classes. On rédigera la réponse suivant le modèle suivant.

```
cat("Test d'ajustement avec la loi gaussien \n")
mu <- mean(...) # moyenne des 200 pots de chocolat
sigma <- sd(...) # écart-type
cat("moyenne : ",mu,"\n")
cat("sigma : ", sigma,"\n")
# Dans la commande suivante, créer 7 valeurs limites
# de 98.5 à 101.5 pour obtenir 6 classes
classes <- seq(...)
facteurs <- cut(poids$chocolats, classes)
nom_classes <- levels(facteurs)
dist_chocolats <- table(facteurs)
print(dist_chocolats)
```

On doit obtenir le tableau suivant

$]98.5, 99]$	$]99, 99.5]$	$]99.5, 100]$	$]100, 100.5]$	$]100.5, 101]$	$]101, 101.5]$
3	20	54	86	26	11

On crée ensuite un vecteur de distribution des 200 poids comme si ces poids suivaient la loi normale $\mathcal{N}(\mu, \sigma)$. On complètera le code

```
dist_normale = vector(mode="numeric", length=6)
dist_normale[1] = pnorm(99,mu,sigma)
dist_normale[2] = ...
dist_normale[3] = ...
dist_normale[4] = ...
dist_normale[5] = ...
dist_normale[6] = 1 - pnorm(101,mu,sigma)
names(dist_normale) <- nom_classes
print(length(poids$chocolats)*dist_normale)
```

On trouvera (après arondi à la première décimale)

$]-\infty, 99]$	$]99, 99.5]$	$]99.5, 100]$	$]100, 100.5]$	$]100.5, 101]$	$]101, +\infty]$
2.1	16.8	55.6	74.8	40.9	9.9

Faire enfin un test d'ajustement du chi-deux

```
test_ajust <- chisq.test(dist_chocolats, p=dist_normale)
print(test_ajust)
```

Refaire ces calculs en utilisant les formules du cours

$$d_{crit} = \sum_{i=1}^{r} \frac{(N_i - np_i^0)^2}{np_i^0}, \quad p_{valeur} = \mathbb{P}(\chi^2 > d_{crit}),$$

où χ^2 est un chi-deux à $r-1$ degrés de liberté. On complètera le code

```
cat("p-valeur du cours : \n")
d_crit <- ...
ddl <- ...
p_valeur <- 1 - pchisq(..., ...)
cat("d_crit : ", d_crit, "\n")
cat("p_valeur : ", p_valeur, "\n")
```

On comparera avec les valeurs données par `chisq.test`.

C Fonction de répartition et quantiles

Lorsqu'on ne dispose pas de logiciel de calculs scientifiques on peut se servir de tables donnant quelques valeurs exactes des fonctions de répartition et des quantiles de distributions classiques.

Loi normale centrée réduite $\mathcal{N}(0,1)$

La loi normale centrée réduite (ou loi de Gauss) est la loi limite d'une distribution d'une somme de variables aléatoires indépendantes de même distribution et renormalisée (d'espérance 0 et de variance 1). Si Z suit une telle loi, on a

TABLE 1 – Loi normale $\mathcal{N}(0,1)$

fonction de répartition :	$\mathbb{P}(Z \leqslant x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{t^2}{2}\Big)\, dt$
paramètres :	$\mu = \mathbb{E}[Z] = 0, \quad \sigma^2 = \mathrm{Var}(Z) = 1$
quantile d'ordre β :	$\mathbb{P}(Z \leqslant q_\beta) = \beta$
dépassement de l'écart absolu :	$\mathbb{P}(\lvert Z\rvert \geqslant z_\alpha) = \alpha$ où $z_\alpha = q_{1-\alpha/2}$

Loi de Student $\mathcal{T}(d)$ ou loi t

Normaliser une variable aléatoire X, c'est se ramener à une variable d'espérance $\mathbb{E}[X] = 0$ et de variance $\mathrm{Var}(X) = 1$ en posant $Z = (X - \mathbb{E}[X])/\sqrt{\mathrm{Var}(X)}$. Dans les problèmes d'intervalle de confiance, on est amené à remplacer $\mathrm{Var}(X)$ par une estimation de celle-ci, c'est-à-dire par une variable V qui suit en général un chi-deux normalisé $V = \chi^2_d/d$. La loi de Student et la loi de la variable

$$T = \frac{U}{\sqrt{V/d}} \quad \text{où} \quad U \overset{\text{loi}}{=} \mathcal{N}(0,1), \quad V \overset{\text{loi}}{=} \chi^2_d$$

et d est le degré de liberté. Si T suit une telle loi, on a

TABLE 2 – Loi de Student $\mathcal{T}(d)$

fonction de répartition :	$\mathbb{P}(T \leqslant x) = \displaystyle\int_{-\infty}^{x} \frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)\sqrt{\pi d}} \Big(1 + \frac{t^2}{d}\Big)^{-\frac{d+1}{2}}\, dt$
paramètres ;	$\mu = \mathbb{E}[T] = 0, \quad \sigma^2 = \mathrm{Var}(T) = d/(d-2)$
quantile d'ordre β :	$\mathbb{P}(T \leqslant q_\beta) = \beta$
dépassement de l'écart absolu :	$\mathbb{P}(\lvert T\rvert \geqslant t_\alpha) = \alpha$ où $t_\alpha = q_{1-\alpha/2}$

Loi du chi-deux $\chi^2(d)$

La loi du chi-deux est la loi d'une somme de carrés de loi normale $\mathcal{N}(0,1)$ et indépendants. Un chi-deux s'apparente au carré d'une distance euclidienne et sert donc souvent dans l'estimation de la dispersion ou de la variabilité d'une distribution. Plus formellement, un chi-deux est donné par

$$\chi^2(d) \overset{\text{loi}}{=} Z_1^2 + \ldots + Z_d^2 \quad \text{où} \quad Z_i \overset{\text{loi}}{=} \mathcal{N}(0,1), \quad \text{les } \{Z_i\} \text{ sont indépendantes}$$

et d est le degré de liberté du système. Si χ^2 est une telle variable, on a

TABLE 3 – Loi du $\chi^2(d)$

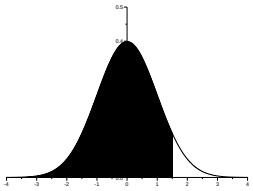
fonction de répartition :	$\mathbb{P}(\chi^2 \leqslant x) = \displaystyle\int_{-\infty}^{x} \frac{\big(\frac{1}{2}\big)^{d/2}}{\Gamma\big(\frac{d}{2}\big)} t^{\frac{d}{2}-1} \exp\Big(-\frac{t}{2}\Big)\, dt$
paramètres ;	$\mu = \mathbb{E}[\chi^2] = 1, \quad \sigma^2 = \mathrm{Var}(\chi^2) = 2d$
quantile d'ordre α :	$\mathbb{P}(\chi^2 \leqslant q_\alpha) = \alpha$

C.1 Loi normale centrée réduite $\mathcal{N}(0,1)$

Table de la fonction de répartition

$$p = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}t^2\big)\, dt$$

Par exemple : si $x = 1.5 + 0.04$ alors $p = 0.9382$



x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

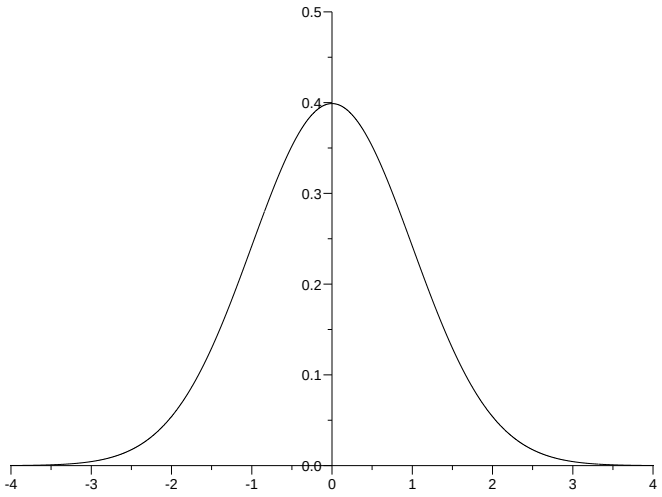
Cas des grandes valeurs de x

x	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7
p	0.998650	0.999032	0.999313	0.999517	0.999663	0.999767	0.999841	0.999892
1-p	0.001350	0.000968	0.000687	0.000483	0.000337	0.000233	0.000159	0.000108

x	3.8	3.9	4.0	4.1	4.2	4.3	4.4	4.5
p	0.999928	0.999952	0.999968	0.999979	0.999987	0.999991	0.999995	0.999997
1-p	0.000072	0.000048	0.000032	0.000021	0.000013	0.000009	0.000005	0.000003

Loi normale centrée réduite : suite

Table de dépassement de l'écart absolu : $\mathbb{P}(|Z| > z_\alpha) = \alpha$
Graphe de la densité $\phi(t) = \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}t^2\big)$.



Par exemple : si $\alpha = 0.1 + 0.03$ alors $z_\alpha = 1.514$.
Cas des grandes valeurs de α :

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695
0.1	1.645	1.598	1.555	1.514	1.476	1.440	1.405	1.372	1.341	1.311
0.2	1.282	1.254	1.227	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.3	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.4	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.5	0.674	0.659	0.643	0.628	0.613	0.598	0.583	0.568	0.553	0.539
0.6	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.426	0.412	0.399
0.7	0.385	0.372	0.358	0.345	0.332	0.319	0.305	0.292	0.279	0.266
0.8	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.9	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013

Cas des petites valeurs de α :

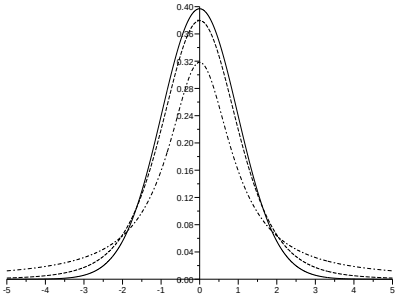
α	0.010	0.005	0.002	0.001	0.0005	0.0002	0.0001	0.00005	0.00002	0.00001
x	2.576	2.807	3.090	3.291	3.481	3.719	3.891	4.056	4.265	4.417

C.2 Loi de Student $\mathcal{T}(d)$

Table de dépassement de l'écart absolu : $\mathbb{P}(|\mathcal{T}| > t_\alpha) = \alpha$
Graphe de la densité $\phi(t) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{\pi d}}(1 + \frac{t^2}{d})^{-\frac{d+1}{2}}$

$\mathcal{T}(d) = \dfrac{U}{\sqrt{V/d}}$ où U et V suivent les lois $N(0,1)$ et $\chi^2(d)$ et sont indépendantes.

$\mathbb{E}(X) = 0$, $\mathrm{Var}(X) = d/(d-2)$.



La première colonne donne le nombre de degrés de liberté *ddl*. La première ligne donne la probabilité α d'être dépassée. Par exemple, si $ddl = 10$ et $\alpha = 0.05$ alors $t_\alpha = 2.228$.

	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

C.3 Loi du chi-deux $\chi^2(d)$

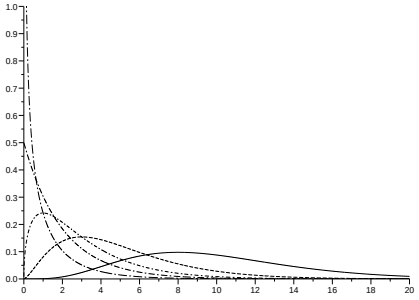
Table des quantiles d'ordre α : $\mathbb{P}(\chi^2(d) < q_\alpha) = \alpha$

La première colonne donne le nombre de degrés de liberté *ddl*. La première ligne donne la probabilité α. Les entrées du tableau donnent q_α. Par exemple, si $ddl = 18$ et $\alpha = 0.01$ alors $q_\alpha = 7.015$.

	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.064	0.455
2	0.002	0.004	0.010	0.020	0.040	0.103	0.211	0.446	1.386
3	0.024	0.039	0.072	0.115	0.185	0.352	0.584	1.005	2.366
4	0.091	0.129	0.207	0.297	0.429	0.711	1.064	1.649	3.357
5	0.210	0.280	0.412	0.554	0.752	1.145	1.610	2.343	4.351
6	0.381	0.486	0.676	0.872	1.134	1.635	2.204	3.070	5.348
7	0.598	0.741	0.989	1.239	1.564	2.167	2.833	3.822	6.346
8	0.857	1.038	1.344	1.646	2.032	2.733	3.490	4.594	7.344
9	1.152	1.370	1.735	2.088	2.532	3.325	4.168	5.380	8.343
10	1.479	1.734	2.156	2.558	3.059	3.940	4.865	6.179	9.342
11	1.834	2.126	2.603	3.053	3.609	4.575	5.578	6.989	10.341
12	2.214	2.543	3.074	3.571	4.178	5.226	6.304	7.807	11.340
13	2.617	2.982	3.565	4.107	4.765	5.892	7.042	8.634	12.340
14	3.041	3.440	4.075	4.660	5.368	6.571	7.790	9.467	13.339
15	3.483	3.916	4.601	5.229	5.985	7.261	8.547	10.307	14.339
16	3.942	4.408	5.142	5.812	6.614	7.962	9.312	11.152	15.338
17	4.416	4.915	5.697	6.408	7.255	8.672	10.085	12.002	16.338
18	4.905	5.436	6.265	7.015	7.906	9.390	10.865	12.857	17.338
19	5.407	5.969	6.844	7.633	8.567	10.117	11.651	13.716	18.338
20	5.921	6.514	7.434	8.260	9.237	10.851	12.443	14.578	19.337
21	6.447	7.070	8.034	8.897	9.915	11.591	13.240	15.445	20.337
22	6.983	7.636	8.643	9.542	10.600	12.338	14.041	16.314	21.337
23	7.529	8.212	9.260	10.196	11.293	13.091	14.848	17.187	22.337
24	8.085	8.796	9.886	10.856	11.992	13.848	15.659	18.062	23.337
25	8.649	9.389	10.520	11.524	12.697	14.611	16.473	18.940	24.337
26	9.222	9.989	11.160	12.198	13.409	15.379	17.292	19.820	25.336
27	9.803	10.597	11.808	12.879	14.125	16.151	18.114	20.703	26.336
28	10.391	11.212	12.461	13.565	14.847	16.928	18.939	21.588	27.336
29	10.986	11.833	13.121	14.256	15.574	17.708	19.768	22.475	28.336
30	11.588	12.461	13.787	14.953	16.306	18.493	20.599	23.364	29.336
35	14.688	15.686	17.192	18.509	20.027	22.465	24.797	27.836	34.336
40	17.916	19.032	20.707	22.164	23.838	26.509	29.051	32.345	39.335
45	21.251	22.477	24.311	25.901	27.720	30.612	33.350	36.884	44.335
50	24.674	26.006	27.991	29.707	31.664	34.764	37.689	41.449	49.335

Loi du chi-deux $\chi^2(d)$: suite

Graphe de la densité $\phi(t) = \frac{(1/2)^{d/2}}{\Gamma(d/2)} t^{d/2-1} \exp(-\frac{t}{2})$



Γ est définie par récurrence par :
$\Gamma(u) = (u-1)\Gamma(u-1)$,
$\Gamma(1) = 1$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

$V = \sum_{i=1}^{d} X_i^2$ où X_i suit la loi $\mathcal{N}(0,1)$.
$\mathbb{P}(V > x) = \int_x^{+\infty} \phi(t)\, dt$

Table des quantiles : suite

	0.75	0.8	0.9	0.95	0.98	0.99	0.995	0.998	0.999
1	1.323	1.642	2.706	3.841	5.412	6.635	7.879	9.550	10.828
2	2.773	3.219	4.605	5.991	7.824	9.210	10.597	12.429	13.816
3	4.108	4.642	6.251	7.815	9.837	11.345	12.838	14.796	16.266
4	5.385	5.989	7.779	9.488	11.668	13.277	14.860	16.924	18.467
5	6.626	7.289	9.236	11.070	13.388	15.086	16.750	18.907	20.515
6	7.841	8.558	10.645	12.592	15.033	16.812	18.548	20.791	22.458
7	9.037	9.803	12.017	14.067	16.622	18.475	20.278	22.601	24.322
8	10.219	11.030	13.362	15.507	18.168	20.090	21.955	24.352	26.124
9	11.389	12.242	14.684	16.919	19.679	21.666	23.589	26.056	27.877
10	12.549	13.442	15.987	18.307	21.161	23.209	25.188	27.722	29.588
11	13.701	14.631	17.275	19.675	22.618	24.725	26.757	29.354	31.264
12	14.845	15.812	18.549	21.026	24.054	26.217	28.300	30.957	32.909
13	15.984	16.985	19.812	22.362	25.472	27.688	29.819	32.535	34.528
14	17.117	18.151	21.064	23.685	26.873	29.141	31.319	34.091	36.123
15	18.245	19.311	22.307	24.996	28.259	30.578	32.801	35.628	37.697
16	19.369	20.465	23.542	26.296	29.633	32.000	34.267	37.146	39.252
17	20.489	21.615	24.769	27.587	30.995	33.409	35.718	38.648	40.790
18	21.605	22.760	25.989	28.869	32.346	34.805	37.156	40.136	42.312
19	22.718	23.900	27.204	30.144	33.687	36.191	38.582	41.610	43.820
20	23.828	25.038	28.412	31.410	35.020	37.566	39.997	43.072	45.315
21	24.935	26.171	29.615	32.671	36.343	38.932	41.401	44.522	46.797
22	26.039	27.301	30.813	33.924	37.659	40.289	42.796	45.962	48.268
23	27.141	28.429	32.007	35.172	38.968	41.638	44.181	47.391	49.728
24	28.241	29.553	33.196	36.415	40.270	42.980	45.559	48.812	51.179
25	29.339	30.675	34.382	37.652	41.566	44.314	46.928	50.223	52.620
26	30.435	31.795	35.563	38.885	42.856	45.642	48.290	51.627	54.052
27	31.528	32.912	36.741	40.113	44.140	46.963	49.645	53.023	55.476
28	32.620	34.027	37.916	41.337	45.419	48.278	50.993	54.411	56.892
29	33.711	35.139	39.087	42.557	46.693	49.588	52.336	55.792	58.301
30	34.800	36.250	40.256	43.773	47.962	50.892	53.672	57.167	59.703
35	40.223	41.778	46.059	49.802	54.244	57.342	60.275	63.955	66.619
40	45.616	47.269	51.805	55.758	60.436	63.691	66.766	70.618	73.402
45	50.985	52.729	57.505	61.656	66.555	69.957	73.166	77.179	80.077
50	56.334	58.164	63.167	67.505	72.613	76.154	79.490	83.657	86.661

C.4 Intervalle de confiance d'une proportion

Estimateurs

— n taille de l'échantillon

— p la proportion à estimer

— $\hat{p}_{\min}$, $\hat{p}_{\max}$, $\hat{p}$ des estimateurs de p

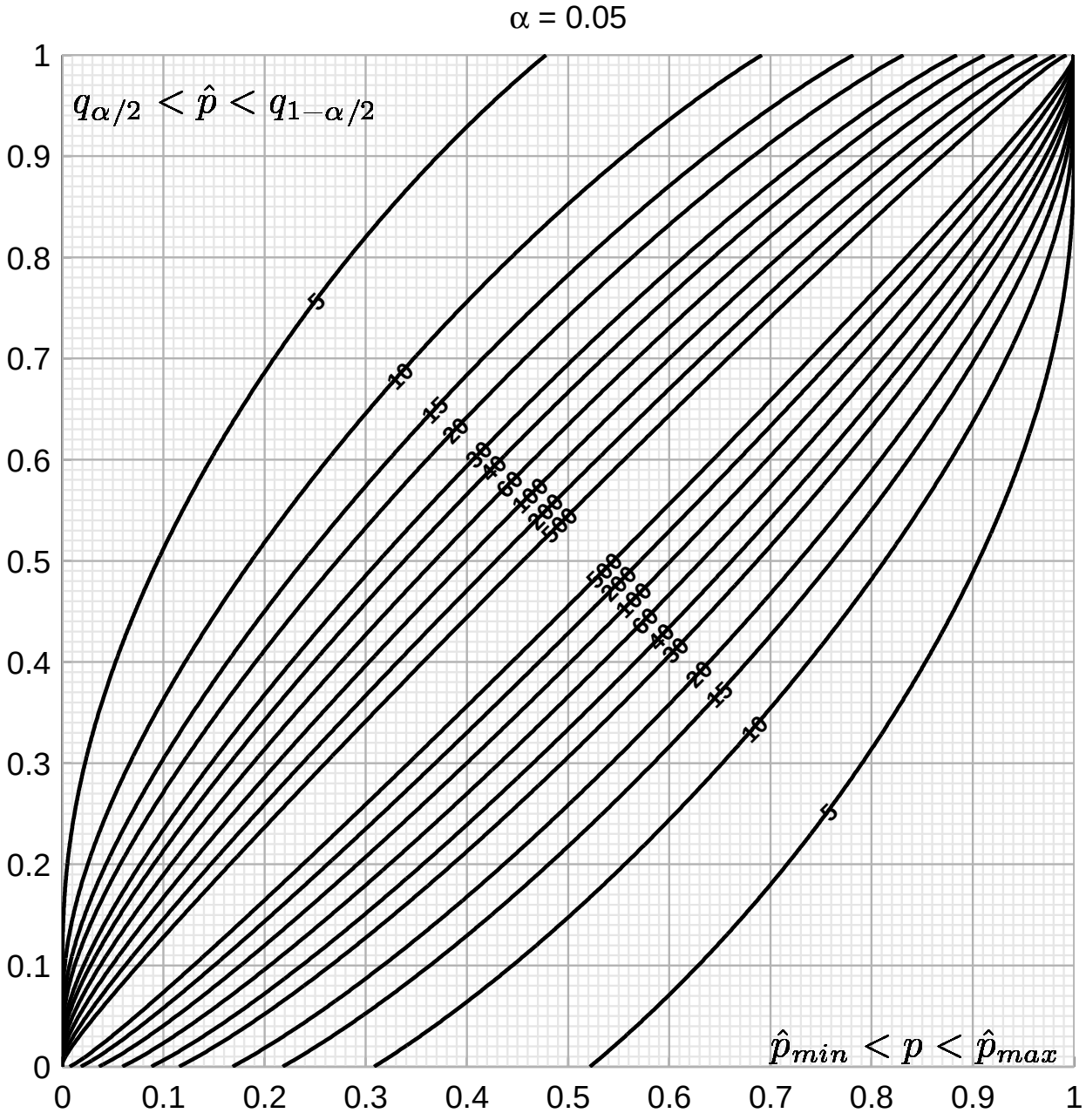
— α risque d'erreur

— $q_{\min}$, $q_{\max}$ les quantiles d'ordre $\frac{1}{2}\alpha$ et $1-\frac{1}{2}\alpha$ de $\hat{p}$

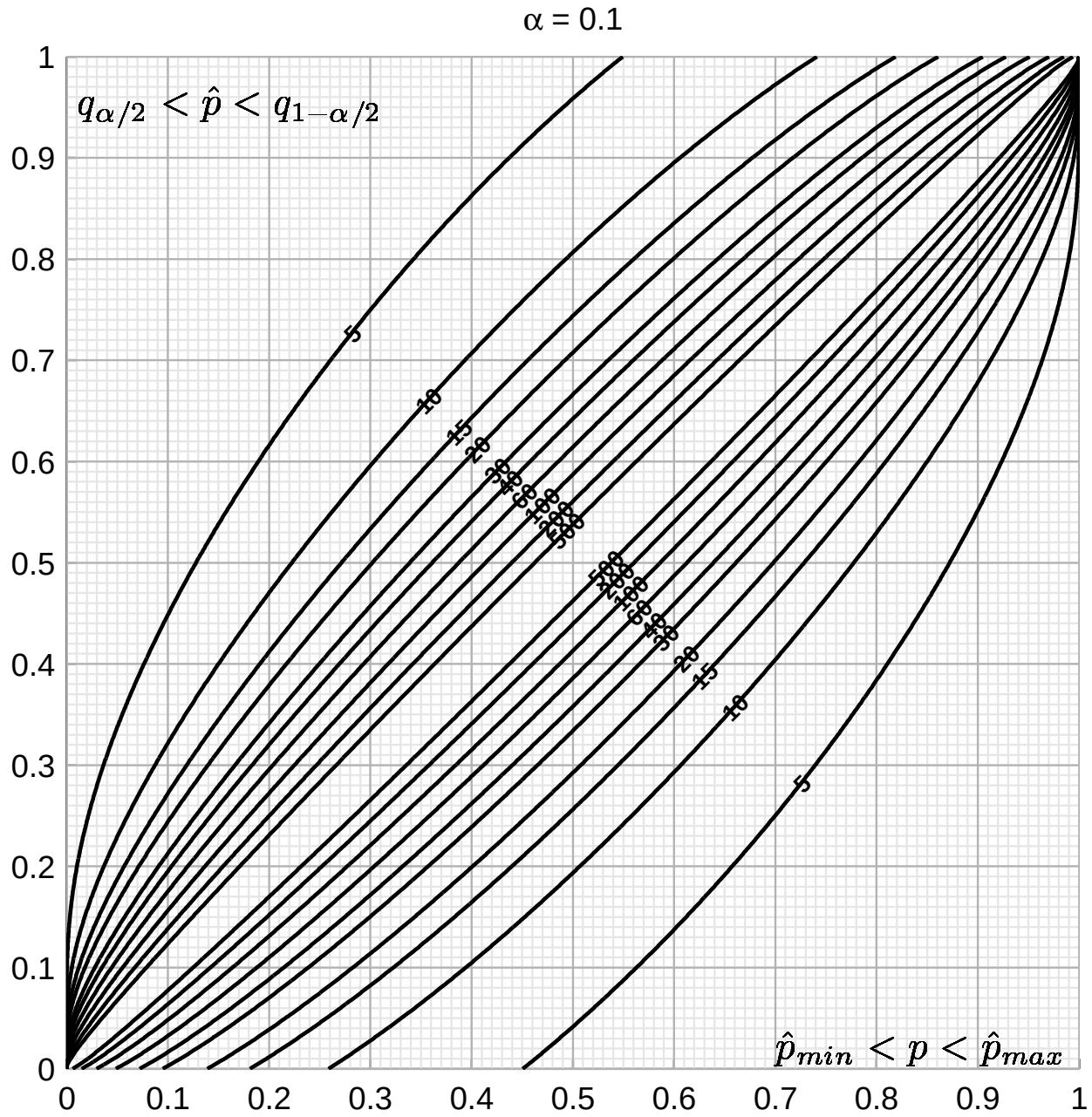
Intervalle de confiance

$$\mathbb{P}(\hat{p}_{\min} \leqslant p \leqslant \hat{p}_{\max}) \geqslant 1-\alpha$$

Tables



Intervalle de confiance d'une proportion : suite



Intervalle de confiance pour des grandes valeurs de n

$$\mathbb{P}\Big(\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqslant p \leqslant \hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\Big) \geqslant 1-\alpha.$$

Calcul de l'erreur $\Delta\hat{\mathbf{p}} = |\hat{\mathbf{p}} + \mathbf{z}_\alpha\sqrt{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})/\mathbf{n}} - \hat{\mathbf{p}}_{\mathbf{max}}|$

	5	10	15	20	30	40	60	100	200	500
0.05	0.522	0.308	0.218	0.168	0.116	0.088	0.060	0.036	0.018	0.007
0.10	0.451	0.259	0.181	0.139	0.095	0.072	0.049	0.030	0.015	0.006

si $n = 15$ et $\alpha = 0.1$ alors $\Delta\hat{p} = 0.181$

D Annales des années précédentes

D.1 Enoncé de l'examen de mars 2012

Exercice 1. On considère une famille à 3 enfants. La probabilité qu'un enfant ait les yeux bleus est de 1/4. La couleur des yeux est indépendante d'un enfant à l'autre.

1. On sait que le premier enfant a les yeux bleus, quelle est la probabilité qu'au moins 2 enfants aient les yeux bleus ?
2. On sait qu'un des enfants a les yeux bleus, quelle est la probabilité qu'au moins 2 enfants aient les yeux bleus ?

Exercice 2. On admet que 50% des ordinateurs personnels utilisent le système d'exploitation Windows, 30% utilisent MacOS et 20% Linux. On sait par ailleurs que 8 utilisateurs Windows sur 10 sont contaminés par un certain virus, 4 utilisateurs MacOS sur 10 le sont et 2 utilisateurs Linux sur 10 le sont. On choisit au hasard un ordinateur personnel et on constate qu'il est infecté par ce virus. Quelle est la probabilité que le système soit sous Windows ?

Exercice 3. On considère dans le tableau suivant la répartition des couples avec enfants en France en 2007 selon le nombre d'enfants (source INSEE)

1 enfant	**2 enfants**	**3 enfants**	**4 enfants**	**5 enfants**
39%	41%	15%	4%	1%

On appelle X la variable aléatoire égale au nombre d'enfants par couple (ayant au moins un enfant). Calculer le nombre moyen d'enfants par couple et son écart type.

Exercice 4. Un tireur à l'arc envoie 10 flèches sur une cible. On admet que chaque tir est indépendant des précédents et que la probabilité d'atteindre la cible est pour chaque tir égale à $p = 0.75$

1. Soit X la variable aléatoire égale au nombre de fois où la cible est atteinte
   (a) Quelle est la loi de probabilité de X ?
   (b) Quelle est l'espérance de X ? Qu'est-ce que cela veut dire concrètement ?
2. On suppose que le tireur parie sur son résultat : il gagne 1€ quand il touche la cible, mais perd $v = 2$€ quand il manque. Soit Y la variable égale au gain du tireur à la fin des 10 tirs.
   (a) Ecrire Y comme une fonction de X.
   (b) Calculer l'espérance de Y.
   (c) Quel montant v faudrait-il utiliser pour que le jeu soit équitable, c'est-à-dire pour que l'espérance du gain soit nulle ?

Exercice 5. On considère la hauteur de 46 pins d'Epinette du Colorado après 20 ans de croissance. Les hauteurs sont mesurées en cm et sont réparties en 6 groupes

hauteur des pins	150–300	300–350	350–450	450–550	550–600	600-650
nombre de pins	2	8	4	17	10	5

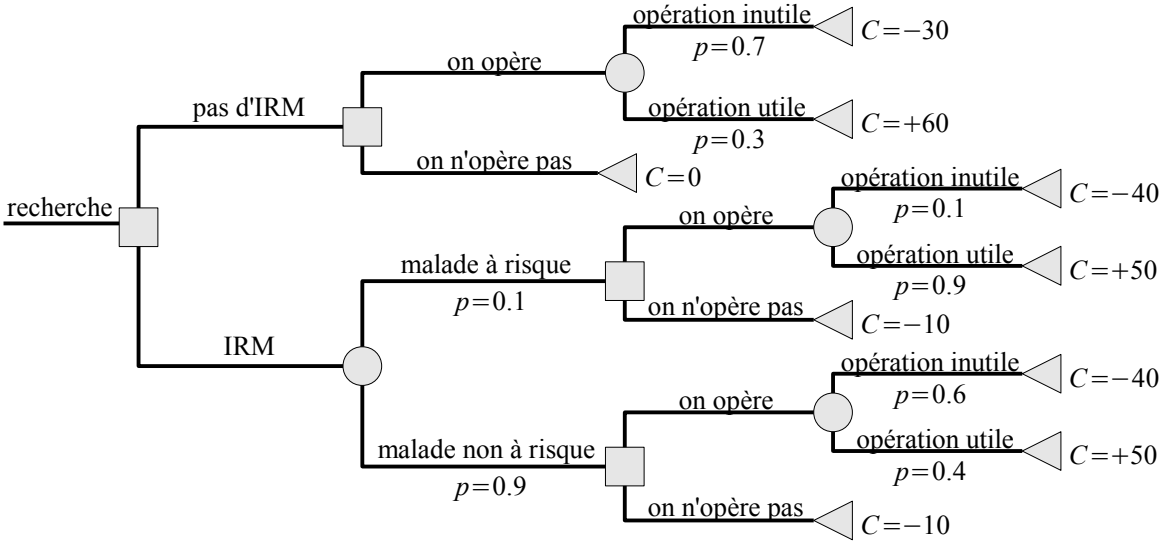
1. Tracer l'histogramme des hauteurs des pins (on remarquera que l'amplitude des classes n'est pas constante). On n'utilisera pas de papier millimétré mais on soignera néanmoins le dessin.
2. Tracer la courbes des effectifs cumulés des pins. Même remarque que précédemment.
3. Déterminer graphiquement la médiane, les deux autres quartiles et résumer vos résultats dans un box-plot.
4. Déterminer la médiane par le calcul.

D.2 Solution de l'examen de mars 2012

D.3 Enoncé de l'examen de juin 2012

Exercice 1. Une secrétaire a perdu un document dans un cabinet formé de 5 bureaux et de l'accueil. La probabilité que le document se trouve à l'accueil est de $1 - p$, celle de se trouver dans un des 5 bureaux est de $p/5$. Après avoir constaté que le document ne se trouvait plus dans les 4 premiers bureaux, déterminer la probabilité qu'il se trouve dans le cinquième.

Exercice 2. On considère l'arbre de décision suivant. Un patient consulte un spécialiste à l'hôpital pour un problème de douleur au genou. Il peut s'agir d'une simple inflammation ou d'une lésion plus sérieuse du ménisque. Pour des raisons d'encombrement des urgences, le médecin peut décider de ne pas réaliser d'IRM. Pour des raisons budgétaires, il peut décider de ne pas opérer par arthroscopie. L'hôpital sait calculer les gains ($C > 0$) ou les perte ($C < 0$) d'une telle opération. L'hôpital connaît aussi le taux des opérations réalisées inutilement (par exemple, la probabilité d'opérer inutilement sans IRM est de $p = 0.7$).



(Constater qu'il n'y a pas de probabilités aux branches de décision).

1. Déterminer la moyenne (ou espérance mathématique) des gains/pertes C lorsqu'on opère sans IRM.
2. Déterminer la moyenne des gains/pertes lorsqu'on opère avec IRM.

Exercice 3. On suppose que la durée de vie d'un certain virus est une variable aléatoire T de densité $f(t)$

$$\begin{cases} f(t) = 0 & \text{pour} \quad t < 0 \quad \text{et} \quad t > 1/\lambda, \\ f(t) = 2\lambda(1 - \lambda t) & \text{pour} \quad 0 \leqslant t \leqslant 1/\lambda, \end{cases}$$

où λ est un paramètre positif.

1. Tracer (sur papier libre) la courbe $y = f(t)$.
2. Déterminer l'espérance et la variance de T.

Exercice 4. Le tableau suivant donne la fréquence relative théorique des voyelles « a, e, i, o, u », avec ou sans accent, dans un texte en français (la somme des fréquences est égale à 1).

a	**e**	**i**	**o**	**u**
0.19	0.37	0.17	0.12	0.15

Le texte suivant est tiré du journal Le Monde électronique (06/06/2012) :

> « Sur Terre, l'évènement est attendu de pied ferme par les astronomes. Vénus - déesse romaine de l'amour, déesse saxonne de la fertilité, déesse maya de la guerre - est une cousine de notre planète, née dans le même nuage de gaz et de poussière il y a 4,6 milliards d'années, et qui partage avec elle le fait d'être enveloppée d'une atmosphère. Or, l'astre a rendez-vous, cette nuit, avec le Soleil. Sa sphère noire se détachera en ombre chinoise devant le disque solaire, dans un parfait alignement, dès l'aube pour la France métropolitaine ».

Un traitement informatisé du texte précédent donne dans le tableau suivant le nombre de chaque voyelle (avec ou sans accent).

a	**e**	**i**	**o**	**u**
38	96	22	19	17

1. Tracer le diagramme en bâton de cette distribution. (C'est inutile de calculer les fréquences).
2. En vous aidant d'un test d'hypothèse dûment commenté, pouvez-vous affirmer au niveau de confiance de 95% que la distribution des voyelles du texte diffère de celle théorique ?
3. Déterminer la probabilité de se tromper en affirmant que les distributions diffèrent.

Exercice 5. On considère dans le tableau suivant la consommation journalière en gramme de Chlorure de sodium NaCl pour 10 sujets de sexe masculin et 5 sujets de sexe féminin.

Homme	2.8	4.1	6.4	7.0	8.4	9.0	10.5	11.4	16.3	20.3
Femme	2.9	5.4	6.7	8.4	13.6					

Déterminer un intervalle de confiance de la différence des moyennes de consommation en NaCl entre hommes et femmes au seuil de confiance de 98%. On séparera bien les formules théoriques du calcul numérique.

D.4 Solution de l'examen de juin 2012

Exercice 1. On découpe l'ensemble de toutes les éventualités en événements disjoints :

$$\Omega = A \cup B1 \cup B2 \cup B3 \cup B4 \cup B5$$

où A désigne l'événement « le document se trouve à l'accueil », $B1$ désigne « le document se trouve dans le bureau 1 »...Si on sait a priori que le document ne se trouve pas dans les quatre premiers bureaux, la probabilité qu'il se trouve dans le cinquième bureau est égale à

$$\mathbb{P}(B5 \mid A \cup B5) = \frac{\mathbb{P}(B5)}{\mathbb{P}(A \cup B5)} = \frac{p/5}{(1-p)+p/5} = \frac{p}{5-4p}.$$

Exercice 2. Dans le modèle proposé de gestion des coûts C, la réalisation d'un IRM coûte $C = -10$, la réalisation d'une opération coûte $C = -30$. De plus l'IRM permet de mieux évaluer la population à risque. Enfin, si l'opération s'avère utile, l'hôpital estime qu'il y a eu un gain de $C = 90$. L'hôpital décide d'opérer tous les patients en choisissant la meilleur stratégie.

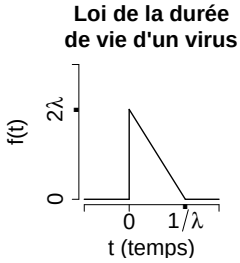
1. Une première stratégie consiste à opérer sans IRM. L'espérance du coût est alors égale à

$$\mathbb{E}[C] = -0.7 * 30 + 0.3 * 60 = -3.$$

2. Une seconde stratégie consiste à opérer avec IRM. L'espérance vaut

$$\mathbb{E}[C] = 0.1 * (-0.1 * 40 + 0.9 * 50) + 0.9 * (-0.6 * 40 + 0.4 * 50) = 0.5.$$

Exercice 4. La durée de vie du virus est modélisée par une loi continue de densité $f(t)$. On constate qu'il s'agit bien d'une densité en vérifiant $\int_0^{1/\lambda} 2\lambda(1-\lambda t)\,dt = 1$.



1. Le tracé de la densité est donné sur la figure D.4 ci-contre.
2. L'espérance de T est donnée par

$$\begin{aligned}\mathbb{E}[T] &= \int_0^{1/\lambda} t\ 2\lambda(1-\lambda t)\,dt \\ &= \frac{1}{\lambda}\int_0^1 2s(1-s)\,ds \\ &= \frac{1}{\lambda}\Big\{\big[s^2\big]_0^1 - \big[\frac{2}{3}s^3\big]_0^1\Big\} = \frac{1}{3\lambda}.\end{aligned}$$

La variance de T est égale à $\mathrm{Var}(T) = \mathbb{E}[T^2] - (\mathbb{E}[T])^2$,

$$\mathbb{E}[T^2] = \int_0^{1/\lambda} t^2\ 2\lambda(1-t\lambda)\,dt = \frac{1}{\lambda^2}\int_0^1 2s^2(1-s)\,ds = \frac{1}{6\lambda^2},$$
$$\mathrm{Var}(T) = \frac{1}{\lambda^2}\Big\{\frac{1}{6} - \frac{1}{9}\Big\} = \frac{1}{18\lambda^2}.$$

Exercice 4.

D.5 Enoncé de d'examen de mars 2013

Exercice 1. Une enquête épidémiologique a été réalisée par la Cellule de l'Institut de Veille Sanitaire, entre le 1er octobre 2010 et le 27 février 2011, pour faire le point sur un cas d'épidémie de rougeole sévissant en région Rhône-Alpes. La répartition des cas par classe d'âge (source Orage) est donnée dans le tableau suivant

Classe d'âge	**Nombre de cas**	**soit en fréquence**
$[0,5[$	423	15.6%
$[5,10[$	289	10.7%
$[10,15[$	565	20.9%
$[15,20[$	645	23.8%
$[20,30[$	529	19.5%
$[30,90[$	255	9.4%
Total	2706	100%

1. Tracer l'histogramme des fréquences des cas de rougeole par classe d'âges. On n'utilisera pas de papier millimétré mais on soignera le dessin.
2. Tracer la courbe des fréquences cumulées. On respectera les mêmes consisgnes que précédemment.
3. Indiquer graphiquement sur le dessin la médiane et les deux autres quartiles ; puis dessiner le boxplot de cette répartition.
4. Donner la formule théorique de la médiane et retrouver par le calcul la valeur qraphique précédente.

Exercice 2. On considère un gène (simplifié) contenant 4 nucléotides $N_1N_2N_3N_4$ que l'on suppose ordonnés. Chacun de ces nucléotides peut prendre les valeurs A, G, T ou C, avec la même probabilité $\frac{1}{4}$ et indépendamment les uns des autres.

Déterminer la probabilité d'obtenir un gène contenant exactement la sequence AA mais pas la séquence AAA ou $AAAA$ (c'est-à-dire un gène de la forme $AAGT$, $TGAA$, $AACA$, ... mais pas $CAAA$ par exemple).

Exercice 3. Une station service est équipée de 6 pompes en état de services et d'une septième qu'on peut éventuellement ouvrir s'il y a trop de monde. Le temps d'attente à chaque pompe en service, exprimé en minutes, est modélisé par une variable T et on admet que $\mathbb{P}(T \leqslant 10) = 75\%$. On suppose que le temps d'attente à une pompe est indépendant de celui des autres pompes. On désigne par X, le nombre de pompes, parmi les 6 en service, dont le temps d'attente est supérieur à 10 mn.

1. Déterminer la loi de la variable X, ainsi que son espérarnce et sa variance.
2. Le gérant de la station service décide d'ouvrir la septième pompe si le temps d'attente, à au moins 4 pompes sur 6, est supérieur à 10 mn. Déterminer la probabilité qu'on mette en marche la septième pompe.

On suppose maintenant que la variable aéatoire suit le loi :

$$\mathbb{P}(T \leqslant t) = \int_0^t \frac{\lambda\, ds}{(\lambda + s)^2}, \qquad \lambda > 0.$$

3. Déterminer λ.
4. Un client a déjà attendu 10 mn à une pompe. Déterminer la probabilité que son temps d'attente ne dépasse pas 15 mn.

Exercice 4. Les gaz non conventionnels se répartissent en trois groupes, gaz de schiste, tight gas et gaz de charbon. De multiples forages ont permis de déterminer avec précision les proportions des différents types de gaz dans le sous-sol. Ces proportions sont données dans le tableau suivant :

Type de gaz	**Gaz de schiste**	**Tight gas**	**Gaz de charbon**
Proportion	47%	29%	24%

Des études sismiques, plus simples à réaliser qu'un forage, permettent de prévoir le type de gisement. La fiabilité du résultat de ces méthodes dépend du type de gaz. Par exemple, pour un gaz de schiste réellement dans le gisement, la méthode sismique prévoit un gaz de schiste 8 fois sur 10, et un tight gas ou un gaz de charbon, chacun, 1 fois sur 10. Le tableau suivant regroupe les différentes prévisions selon le type de gaz :

		Gas réellement dans le gisement		
		Gas de schiste	**Tight gas**	**Gaz de charbon**
Gaz prévu par relevé sismique	**Gaz de schiste**	50%	20%	20%
	Tight gaz	25%	60%	20%
	Gaz de charbon	25%	20%	60%

1. On arrive sur un gisement inconnu et on réalise un relevé sismique. Quelle est la probabilité d'avoir du gaz de schiste par relevé sismique ?
2. Une étude est réalisée dans la province de Neuquen en Patagonie. Les relevés par la méthode sismique indiquent un gisement de gaz de schiste. Quelle est la probabilité que ce gaz soit réellement du gaz de schiste ?

D.6 Solution de l'examen de mars 2013

Exercice 1.

1. Le seul point important à respecter dans le tracé de l'histogramme des fréquences est de disposer des rectangles, au dessus de chaque classe, de hauteur inversement proportionnelle à la largeur de la classe. On choisit d'abord une unité de base : soit une classe d'âge de 5 ans ; puis on répartit la population par unité de classe d'âge. Par exemple, 19.5 % cas de rougeole ont concerné des personnes de 20 à 30 ans : soit 19.5/2 % de cas pour la classe $[20, 25[$ et 19.5/2 % de cas pour la classe $[25, 30[$. L'histogramme est représenté sur la figure 8.
2. Les fréquences cumulées par classe en % sont données dans le tableau suivant

Classe	$[0, 5[$	$[5, 10[$	$[10, 15[$	$15, 20$	$[20, 30[$	$[30, 90[$
Fréquences en %	15.6	26.3	47.2	71.0	90.6	100.0

Dans le cas du tracé de la courbe des fréquences cumulées, il n'est pas nécessaire de tenir compte de la largeur des classes. La courbe est donnée sur la figure 8.

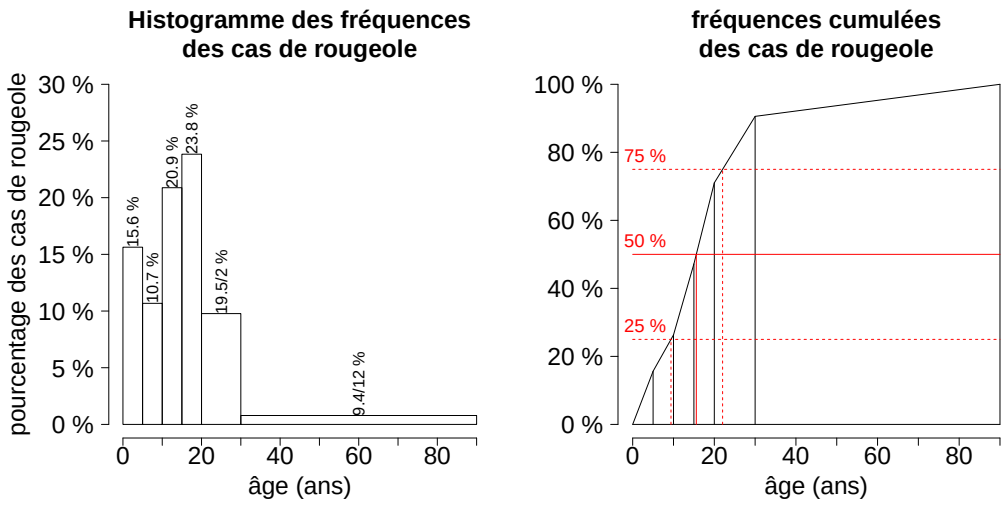


FIGURE 8 – Répartition des cas de rougeole entre le 01/10/2010 et le 27/02/2011 en région Rhône-Alpes.

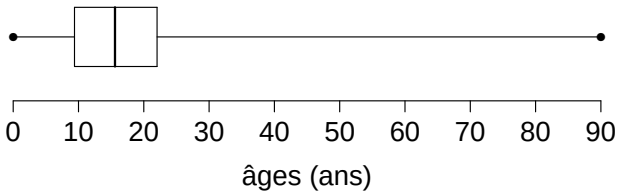


FIGURE 9 – Répartition en quartiles, médianes et valeurs extrêmes des âges des cas de rougeole.

3. Les différents quartiles sont obtenus graphiquement sur l'axe horizontale des âges de la courbes des fréquences cumulées. Le box plot résume dans un graphique, le premier et troisième quartiles, la médianes et les deux valeurs extrêmes des âges. On obtient la figure 9.
4. La classe médiane est de $[15, 20[$ ans. Elle correspond à une proportion des cas de rougeole située dans l'intervalle $[47.2\ \%, 71.0\ \%[$ La médiane est obtenue en reportant 50 % de manière linéaire dans la calsse médaine, soit

$$\text{médiane} = 15 + (20 - 15) * (50 - 47.2)/(71.0 - 47.2) = 15.6 \text{ ans.}$$

Les deux autres quartiles sont $q_{25\%} = 9.4$ ans et $q_{75\%} = 22.0$ ans.

Exercices 2. On note par Ω l'espace de toutes les séquences de 4 nucléotides. Chacun de ces nucléotides apparait avec la même probabilité ; la probabilité d'observer un nucléotide en particulier est donc égale à $1/\text{Card}(\Omega) = 1/4^4$. On considère l'ensemble E des séquences de 4 nucléotides contenant AA mais pas AAA ou $AAAA$. La probabilité d'observer un tel ensemble est $\mathbb{P}(E) = \text{Card}(E)/\text{Card}(\Omega)$. On décrit E comme la réunion disjointe de 2 sous ensembles

$$E = E_1 \cup E_2 = \{\ AA*A\ ,\ A*AA\ \} \cup \{\ AA**\ ,\ *AA*\ ,\ **AA\ \}$$

où $*$ désigne un nucléotide quelconque parmi G, T, C. Comme

$$\text{Card}(E_1) = 2 * 3, \quad \text{Card}(E_2) = 3 * 3^2, \quad \text{Card}(E) = \text{Card}(E_1) + \text{Card}(E_2) = 33,$$

la probabilité d'observer E est donc $\mathbb{P}(E) = 33/256 = 12.9\ \%$.

Exercice 3.

1. X suit une loi binomiale de paramètre $\mathcal{B}(n = 6,\ p = 25\ \%)$, où p désigne la probabilité que le temps d'attente à une pompe est supérieur à 10 mn. L'espérance et la variance sont données par

$$\mathbb{E}[X] = np = 1.5 \text{ (pompes)}, \quad \mathrm{Var}(X) = np(1-p) = 1.125 \text{ (pompes}^2\text{)}.$$

2. On désigne par E l'événement « Le gérant ouvre la septième pompe ». On a

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(X \geqslant 4)\\ &= \mathbb{P}(X=4) + \mathbb{P}(X=5) + \mathbb{P}(X=6)\\ &= \binom{6}{4} 0.25^4 * 0.75^2 + \binom{6}{5} 0.25^5 * 0.75 + 0.25^6\\ &= 15 * 0.25^4 * 0.75^2 + 6 * 0.25^5 * 0.75 + 0.25^6 = 3.8\ \%.\end{aligned}$$

3. On constate bien que la formule donnée dans l'énoncé est une fonction de répartition : $\int_0^{+\infty} \lambda/(\lambda+s)^2\, ds = 1$, quelque soit la valeur de $\lambda > 0$. Par hypothèse, $\mathbb{P}(T \leqslant 10) = 0.75$; comme

$$\mathbb{P}(T \leqslant t) = \Big[-\frac{\lambda}{\lambda+s} \Big]_{s=0}^{s=t} = 1 - \frac{\lambda}{\lambda+t} = \frac{t}{\lambda+t}, \quad \text{on a}$$
$$\frac{10}{\lambda+10} = 0.75 \quad \Longleftrightarrow \quad \lambda = 10 * \frac{0.25}{0.75} = 3.33 \text{ mn}.$$

4. Si l'événement $T > 10$ mn est réalisé, la probabilité que $T \leqslant 15$ mn est donnée par

$$\mathbb{P}(T \leqslant 15 \mid T > 10) = \frac{\mathbb{P}(10 < T \leqslant 15)}{\mathbb{P}(T > 10)}$$

Comme

$$\mathbb{P}(10 < T \leqslant 15) = \int_{10}^{15} \frac{\lambda\, ds}{(\lambda+s)^2} = \Big[-\frac{\lambda}{\lambda+s} \Big]_{s=10}^{s=15} = \frac{5\lambda}{(\lambda+10)(\lambda+15)},$$
$$\mathbb{P}(T > 10) = \int_{10}^{+\infty} \frac{\lambda\, ds}{(\lambda+s)^2} = \Big[-\frac{\lambda}{\lambda+s} \Big]_{s=10}^{s=+\infty} = \frac{\lambda}{\lambda+s},$$

on obtient $\mathbb{P}(T \leqslant 15 \mid T < 10) = 5/(\lambda+15) = 5/(3.33+15) = 27.3\ \%$.

Exercice 4. L'énoncé suggère deux méthodes pour déterminer le type de gaz se trouvant dans le gisement : une méthode sûre et onéreuse par forage, et une méthode moins précise et moins couteuse par relevé sismique. Le premier tableau donne la proportion exacte des 3 types de gaz dans le sous-sol qu'on a pu obtenir avec précision à la suite de tous les forages déjà réalisés. Le deuxième tableau donne (en colonne) la proportion des 3 type de gaz après relevé sismique lorsqu'on connait à l'avance ce que le sous-sol recèle. Ce type de tableau est réalisé avant commercialisation de l'appareil et sert détalonnage par la suite.

1. On est en présence de deux types d'informations ou deux variables aléatoires : la variable $X =$ type de gaz réel obtenu par forage et la variable $Y =$ type de gaz prévu par relevé sismique. Le premier tableau donne la loi de X. Le deuxième tableau est un tableau de probabilités conditionnelles

$$\mathbb{P}(Y = \text{gaz prévu} \mid X = \text{gaz réel}) = \mathbb{P}(Y = GP \mid X = GR).$$

Pour calculer les proportions de gaz prévu par relevé sismique après forage, soit $\mathbb{P}(Y = GP \text{ et } X = GR)$, le plus simple est de construire un arbre de probabilité comme dans la figure 10. Le calcul utilise la formule

$$\mathbb{P}(Y = GP \text{ et } X = GR) = \mathbb{P}(Y = GP \mid X = GR)\ \mathbb{P}(X = GR).$$

La probabilité d'obtenir du gaz de schiste par relevé sismique $\mathbb{P}(Y = \text{Schiste})$ (sans connaître à

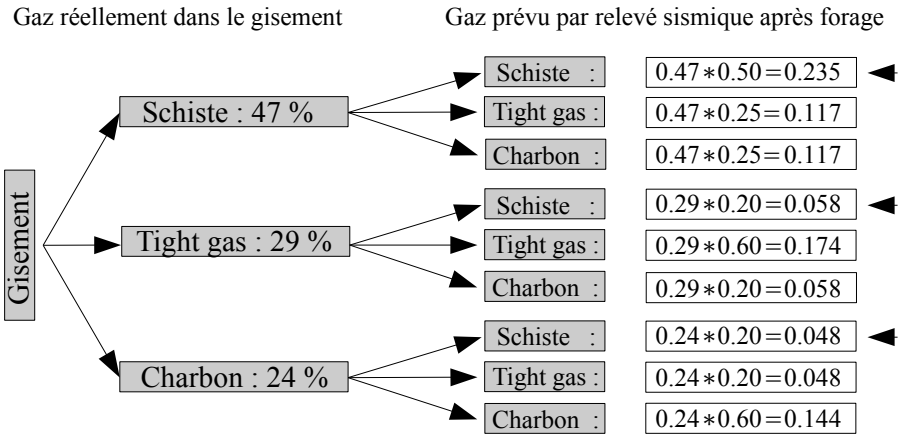


FIGURE 10 – Arbre de probabilités conditionnelles.

l'avance le type de gaz réel) est alors obtenu par la calcul suivant

$$\begin{aligned}\mathbb{P}(Y = \text{Schiste}) &= \mathbb{P}(Y = \text{Schiste et } X = \text{Schiste}) \\ &\quad + \mathbb{P}(Y = \text{Schiste et } X = \text{Tight}) \\ &\quad + \mathbb{P}(Y = \text{Schiste et } X = \text{Charbon}) \\ &= 0.235 + 0.058 + 0.048 = 34.1\ \%.\end{aligned}$$

On aurait pu calculer la probabilité d'obtenir un des 3 gaz par relevé sismique et on aurait ainsi obtenu la loi de Y :

Type de gaz	**Gaz de schiste**	Tight gas	Gaz de charbon
Proportion	34.1 %	33.95 %	31.95 %

On obtient sensiblement la même proportion d'obtenir un des 3 gaz par relevé sismique.

2. Connaître le type de gaz réel lorsqu'on vient d'obtenir le type de gaz par relevé sismique, c'est calculer $\mathbb{P}(X = GR \mid Y = GP)$. Si le relevé donne du schiste, la probabilité que le gisement contienne réellement du schiste est donc

$$\begin{aligned}\mathbb{P}(X = \text{Schiste} \mid Y = \text{Schiste}) &= \frac{\mathbb{P}(X = \text{Schiste et } Y = \text{Schiste})}{\mathbb{P}(Y = \text{Schiste})} \\ &= \frac{0.235}{0.341} = 68.9\ \%.\end{aligned}$$

De même on aurait pu calculer

$$\begin{aligned}\mathbb{P}(X = \text{Tight} \mid Y = \text{Tight}) &= 51.3\ \% \quad \text{et} \\ \mathbb{P}(X = \text{Charbon} \mid Y = \text{Charbon}) &= 45.1\ \%.\end{aligned}$$

On constate que l'appareil de mesure est plus précis s'il s'agit de détecter du gaz de schiste.

D.7 Enoncé de l'examen de juin 2013

Exercice 1. Dans la population française, les groupes sanguins se répartissent de la façon suivante :

O	**A**	**B**	**AB**
43%	45%	9%	3%

Lors d'une transfusion sanguine, un donneur est compatible avec un receveur du groupe A si et seulement si le donneur est du groupe O ou du groupe A.

1. On choisit un donneur français au hasard pour un receveur du groupe A. Quelle est la probabilité que le donneur soit compatible ?

2. On effectue la transfusion à ce receveur du groupe A et elle se passe mal. Quelle est la probabilité que le donneur soit du groupe B ?

   Voici le tableau des compatibilités donneur/receveur (les couples compatibles sont marqués par des croix).

receveur \ **donneur**	**O**	**A**	**B**	**AB**
O	x			
A	x	x		
B	x		x	
AB	x	x	x	x

3. On choisit un donneur et un receveur français au hasard et de façon indépendante. Quelle est la probabilité qu'ils soient compatibles ?

Exercice 2. On veut estimer la proportion d'individus de groupe sanguin O dans la population allemande. On prélève un échantillon de 92 personnes. Parmi ceux-ci, 47 sont du groupe O. Donner un intervalle de confiance, au risque de 5%, de la proportion d'individus du groupe O dans la population totale.

Exercice 3. Un cultivateur plante chaque année des pommes de terre sur un terrain T. Les années paires, il plante des Bintjes, les années impaires il plante des Charlotte. Voici les rendements annuels (en tonnes par hectare) de la Bintje, de l'année 2004 à l'année 2012 :

Rendement annuel des Bintjes					
année	2004	2006	2008	2010	2012
rendement (en t/ha)	36	55	47	58	52

1. On suppose que ces 5 années de culture de Bintje constituent un échantillon représentatif. Donner un intervalle de confiance au risque de 5% du rendement annuel moyen de la Bintje sur le terrain T.
2. Voici les rendements annuels (en tonnes par hectare) de la Charlotte, de l'année 2003 à l'année 2011 :

Rendement annuel des Charlottes					
année	2003	2005	2007	2009	2011
rendement (en t/ha)	45	38	43	41	38

   On suppose que ces 5 années de culture de Charlotte constituent également un échantillon représentatif. Pourquoi ne peut-on pas dire que les deux échantillons ci-dessus sont appariés ?
3. Au niveau de confiance de 90%, peut-on dire que le rendement moyen de la Bintje sur le terrain T est différent du rendement moyen de la Charlotte ? On répondra par un test de comparaison de moyennes.
4. Donner la p-valeur de ce test.

Exercice 4. On cherche à déterminer si l'espérance de vie à 70 ans dépend de la corpulence. On suit un échantillon de 80 personnes sur 20 ans. Au début de l'étude, ces 80 personnes ont 70 ans. A l'issue de l'étude, les effectifs de chaque classe de corpulence et d'âge de décès sont synthétisés dans le tableau suivant :

corpulence \ **âge du décès**	$[\mathbf{70};\mathbf{80}[$	$[\mathbf{80};\mathbf{90}[$	$[\mathbf{90};\infty[$
poids normal	10	7	4
en surpoids	6	24	5
obèse	3	14	7

Au risque de $\alpha = 1\%$, peut-on dire que âge de décès et corpulence sont indépendants ?

D.8 Solution de l'examen de juin 2013

Exercice 1.

1. Il y a 4 groupes sanguin possibles $\Omega = \{O, A, B, AB\}$. On s'intéresse à l'événement

$$E = \text{« la transfusion est compatible avec un receveur du groupe } A \text{ »} = O \cup A.$$

   D'où la probabilité d'être compatible $\mathbb{P}(E) = \mathbb{P}(O) + \mathbb{P}(A) = 43\% + 45\% = 88\%$.

2. On définit l'événement $F =$ « la transfusion n'est pas compatible » $= \Omega \backslash E = B \cup AB$. La probabilité que le donneur soit B lorsqu'on sait que la transfusion est incompatble est donc

$$\mathbb{P}(B \mid F) = \frac{\mathbb{P}(B \cup F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(B)}{\mathbb{P}(F)} = \frac{9}{9+3} = 75\%.$$

3. On considère l'ensemble de tous les choix possibles $(D, R) =$ (donneur, receveur)

$$\Omega = \{(O,O), (O,A), (O,B), (O,AB), (A,O), (A,A), (A,B), \ldots, (AB,AB)\}.$$

   On s'intéresse à l'événement $E =$ « (D, R) sont compatibles ». Par hypothèse, la probabilité d'un choix particulier de (donneur, receveur) est donnée par $\mathbb{P}(D, R) = \mathbb{P}(D)\mathbb{P}(R)$. Par exemple $\mathbb{P}(O, A) = 43\% * 45\%$. D'où la probabilité d'être compatible lorsqu'on choit un donneur et un receveur au hasard

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(O)[\mathbb{P}(O) + \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(AB)] \\ &\quad + \mathbb{P}(A)[\mathbb{P}(A) + \mathbb{P}(AB)] + \mathbb{P}(B)[\mathbb{P}(B) + \mathbb{P}(AB)] + \mathbb{P}(AB)\mathbb{P}(AB) \\ &= 43\% + 45\%48\% + 9\%12\% + 3\%3\% = 0.6577.\end{aligned}$$

Exercice 2. L'intervalle de confiance de la proportion p d'être du groupe sanguin O est

$$p = p_{obs} \pm z_\alpha \sqrt{p_{obs}(1 - p_{obs})/n}.$$

Numériquement on obtient : $n = 92$, $p_{obs} = 47/92$, $\alpha = 5\%$, $z_\alpha = 1.96$, d'où

$$P = 0.51 \pm 0.11 \quad \text{ou bien} \quad 0.40 < p < 0.62.$$

(Le code suivant du logiciel `R`,

```
test.gs <- prop.test(47,92); ci <- test.gs$conf.int; ci
```

donne directement le résultat `0.4051209 0.6157014`).

Exercice 3.

1. L'intervalle de confiance de la moyenne μ_B des rendements annuels de la Bintje est donné par

$$\mu_B = \bar{x} \pm t_\alpha s_B/\sqrt{n} \quad \text{avec} \quad \bar{x} = \sum\nolimits_i x_i/n \quad \text{et} \quad s_B = \sum\nolimits_i (x_i - \bar{x})^2/(n-1).$$

   Numériquement on obtient : $n = 5$, $\bar{x}_B = 49.6$ t/ha, $\alpha = 0.05$, $t_\alpha = 2.776$, $s_B = 8.62$ t/ha et donc

$$\mu_B \in [38.89,\ 60.31] \text{ t/ha.}$$

   (En utilisant la commande

```
test.binjte <- t.test(bintje); ci <- test.binjte$conf.int; ci
```

   on obtiendrait avec le logiciel R : $\mu_B \in [38.89717,\ 60.30283]$)

2. Les conditions de culture d'une année à l'autre sont des variables aléatoires indépendantes. On est donc en présence de 10 « individus » que l'on sépare en 2 groupes. Ces 2 groupes restent indépendants et les 2 échantillons ne sont donc pas appariés.

3. On note μ_C le rendement annuel théorique de la Charlotte. L'hypothèse nulle est

$$H_0 = \text{« les deux rendements sont égaux »} = \{\mu_B = \mu_C\}.$$

Le critère de rejet est donné par la variable de Student

$$T_{\text{rejet}} := \frac{|\bar{X}_B - \bar{X}_C|}{S_{BC}\sqrt{1/n_B + 1/n_C}} > t_\alpha, \quad S_{BC} = \sqrt{\frac{(n_B-1)S_B^2 + (n_C-1)S_C^2}{n_B + n_C - 2}}$$

Numériquement, $\bar{x}_C = 41$ t/ha, $s_C = 3.08$ t/ha, $s_{BC} = 6.47$ t/ha, $ddl = 8$, $\alpha = 0.1$, $t_\alpha = 1.860$, $t_{\text{rejet}} = 2.1$. On vérifie bien que $t_{\text{rejet}} > t_\alpha$: on rejette H_0, le rendement de la Charlotte est différent de celui de la Bintje.

4. La p-valeur est l'erreur statistique que l'on commet en rejetant à tort H_0. C'est l'erreur α qui vérifie

$$t_\alpha = \frac{|\bar{x}_B - \bar{x}_C|}{s_{BC}\sqrt{1/n_B + 1/n_C}} \quad \Longleftrightarrow \quad \alpha = \mathbb{P}\Big(|\mathcal{T}_8| > \frac{|\bar{x}_B - \bar{x}_C|}{s_{BC}\sqrt{1/n_B + 1/n_C}}\Big).$$

Les tables proposées ne donnent qu'un encadrement de la variable de rejet $2.1 \in [1.860, 2.306]$; on obtient seulement p-valeur $\in [0.05,\ 0.1]$. La commande de `R`

```
t.test(bintje, charlotte, alternative = "two.sided",
                                          paired=FALSE, var.equal = TRUE)
```

donne la valeur exacte de la p-valeur `0.06886`, qu'on aurait pu obtenir aussi avec la commande `p.val <- 2-2*pt(t.rejet,8);p.val`. Les formules du cours suppose que les variances sont égales ; sans cette hypothèse et avec le commande

```
t.test(bintje, charlotte, alternative = "two.sided",
                                        paired=FALSE, var.equal = FALSE)
```

on obtient le résultat : `t = 2.1007, df = 5.006, p-value = 0.0896`. On obtient une probabilité de se tromper un peu plus grande, qui se justifie par le fait que l'on dispose de moins d'information initialement.

Exercice 4. On s'agit ici d'un test d'indépendance entre les variables « poids » et « âge du décès ». L'hypothèse nulle toujours l'indépendance des deux variables que l'on considère. La critère de rejet est donnée par la formule

$$D^2 := \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - N_{*j}N_{i*}/n)^2}{N_{*j}N_{i*}/n} > q_{1-\alpha}$$

où q_α est le quantile d'ordre α d'un chi-deux à $(r-1)(s-1)$ degrés de liberté. Numériquement, il est commode de disposer les calculs dans une table donnant à la fois les effectifs observés et les effectifs théoriques

	$\mathbf{70-80}$	$\mathbf{80-90}$	$\mathbf{90-\infty}$	$\mathbf{N_{i*}}$
normal	10 / 4.99	7 / 11.81	4 / 4.20	21
surpoids	6 / 8.31	24 / 19.69	5 / 7.00	35
obèse	3 / 5.70	14 / 13.50	7 / 4.80	24
$\mathbf{N_{*j}}$	19	45	16	80

Numériquement, on trouve la valeur de rejet $d^2 = 11.473$. Pour $\alpha = 0.01$, pour un degré de liberté $ddl = 4$, le quantile du chi-deux vaut $q_{1-\alpha} = 13.267$. Comme $11.473 < 13.267$, on accepte H_0 (ou plutôt, on ne rejette pas H_0 par insuffisance de preuve). La commande `R`

```
chisq.test(tableau)$p.value
```

donne la p-valeur $p_{val} = 0.022$: au risque 2.2% on peut rejeter H_0.

D.9 Enoncé de d'examen de mars 2014

Exercice 1. Des vents d'une violence exceptionnelle ont traversé le nord de la France le 26 décembre 1999 du Finistère à Strasbourg. Le tableau suivant fournit pour 75 stations les vitesses moyennes mesurées sur les rafales :

Plages de vitesses en km/h	$]70,90]$	$]90,110]$	$]110,130]$	$]130,150]$	$]150,170]$
Nombre de stations	8	8	25	22	12

1. Tracer la courbe des effectifs cumulés.
2. Placer les 3 quartiles sur la figure et tracer le box plot.
3. Déterminer la médiane par le calcul.

Exercice 2. Le tableau suivant fournit le nombre de décès accidentels en 2005 en France selon l'âge et certaines causes d'accidents.

	Type des accidents			
Age (ans)	**Tous types** (dont voiture, chute, noyade)	**voiture**	**chute**	**noyade**
moins de 5	541	163	8	113
5-14	595	354	7	75
15-24	2822	2112	47	129
25-34	2353	1376	60	83
35-44	3082	1385	121	96
45-54	2455	1072	174	70
55-64	1501	701	189	43
65-74	1539	607	332	35
plus de 75	4687	897	1722	48
Total	19575	8667	2660	692

1. On choisit au hasard une personne morte accidentellement en 2005 en France. Dans chacun des cas suivants, déterminer la probabilité que
   (a) l'âge de cette personne soit au moins de 15 ans,
   (b) le décès fait suite à un accident de voiture,
   (c) le décès fait suite à un accident de voiture sachant que l'âge de la personne était compris entre 25 et 44 ans,
   (d) le décès est une noyade sachant qu'il ne s'agissait pas d'un accident de voiture et que l'âge de la personne était inférieur à 34 ans.
2. On choisit maintenant au hasard une personne dans la population française. Le tableau précédent permet-il de calculer la probabilité que cette personne décède suite à un accident de voiture ?

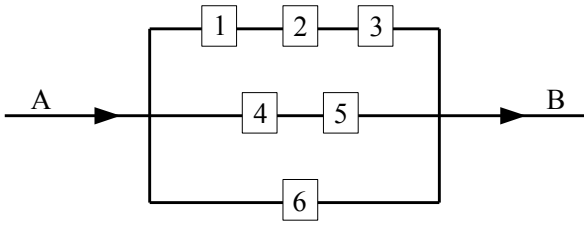


FIGURE 11 – Système de canalisations avec 3 valves.

Exercice 3. On considère un système de canalisations d'eau s'écoulant du point A vers le point B à travers 6 valves numérotées de 1 à 6, comme c'est indiqué sur la figure 11. Un signal est envoyé pour ouvrir les valves simultanément. Les valves sont cependant un peu défectueuses et s'ouvrent 9 fois sur 10. On admet qu'elles fonctionnent de manière indépendante les unes des autres. Une canalisation est dite ouverte si toutes les valves se situant sur le chemin sont ouvertes. On note X, le nombre de canalisations ouvertes permettant de laisser écouler l'eau de A vers B : $X = 0, 1, 2, 3$.

1. Déterminer la probabilité que les valves 1, 2 et 3 soient ouvertes simultanément.
2. Déterminer la loi de X.
3. Déterminer la probabilité que l'eau s'écoule de A vers B.

Exercice 4. Une machine fabrique des roulements à bille de diamètre variable selon une loi normale de paramètres $\mu = 1.005$ mm et $\sigma = 0.010$ mm. Les spécifications techniques fournies par le fabriquant affirment que le diamètre se situe dans l'intervalle 1.000 ± 0.020 mm. Un contrôle technique en fin de chaîne écarte les roulements à bille non conformes.

1. On prend un roulement à bille au hasard sur la chaîne de production, quelle est la probabilité qu'il ne soit pas conforme.
2. Sans faire aucun calcul et en faisant appel à l'intuition, quelle valeur du diamètre moyen μ pourrait-on choisir de façon que la probabilité calculée précédemment soit la plus petite possible ?

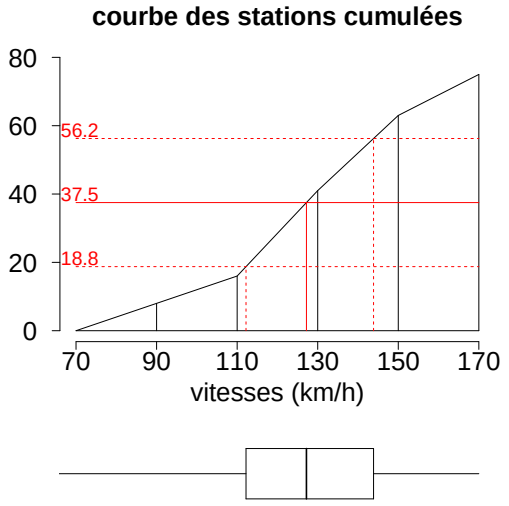
Exercice 5. (Plus difficile) On considère le jeu de dé suivant : la mise est de 100 €; le joueur lance un dé parfait tant que la face du dé est égale à 1; il ne peut jouer que 4 fois au maximum; si au premier coup, la face du dé est différente de 1, le joueur gagne 80 € et le jeu s'arrête; sinon il rejoue une deuxième fois, et si au deuxième coup, la face du dé est différente de 1, le joueur gagne $3 * 80$ €; à chaque fois qu'il rejoue et que la face du dé est différente de 1, le gain est triplé; si la face 1 apparait aux 4 lancés, le joueur perd sa mise et le jeu s'arrête.

On note X le nombre de fois que la face 1 apparait dans ce jeu, (le jeu s'arrête dès que la face est différente de 1). On note Y le bénéfice en fin de jeu, c'est-à-dire, le gain obtenu du joueur privé de la somme mise au début, soit $Y = 80 * 3^X - 100$ €, si $X \leqslant 3$, et $Y = -100$ €, si $X = 4$.

1. Déterminer la loi de X.
2. Déterminer l'espérance du bénéfice.

D.10 Solution de l'examen de mars 2014

Exercice 1.



– Les 3 quartiles valent approximativement sur la figure

$q_{25\%}$	$q_{50\%}$	$q_{75\%}$
112	127	144

– La classe médiane est $]110, 130]$, les effectifs cumulés le long de cette classe sont 16 et 41.
– Le calcule théorique de la médiane donne

$$q_{50\%} = 110 + 20 \times \frac{75/2 - 16}{41 - 16} = 127.2.$$

FIGURE 12 – Répartion des vitesses extrêmes du vent.

Exercice 2.

1. (a) $\mathbb{P}(\text{âge} \geqslant 15) = \dfrac{\text{effectif d'âge } \geqslant 15}{\text{effectif total}} = \dfrac{19575 - 541 - 595}{19575} = \dfrac{18439}{19575} \simeq \ldots\ldots\ldots\ldots\ldots 0.942.$

   (b) $\mathbb{P}(\text{accident\_de\_voiture}) = \dfrac{8667}{19575} \simeq \ldots\ldots\ldots\ldots\ldots 0.443.$

   (c) $\mathbb{P}(\text{accident\_de\_voiture} \mid 25 \leqslant \text{âge} \leqslant 44)$
   $= \dfrac{\mathbb{P}(\text{accident\_de\_voiture ET } (25 \leqslant \text{âge} \leqslant 44))}{\mathbb{P}(25 \leqslant \text{âge} \leqslant 44)} = \dfrac{1376 + 1385}{2353 + 3082} \simeq \ldots\ldots\ldots\ldots\ldots 0.508.$

(d) $\mathbb{P}(\text{accident\_par\_noyade} \mid \text{accident\_hors\_voiture ET } (\text{âge} \leqslant 34))$
$$= \frac{\text{accident\_par\_noyade ET } (\text{âge} \leqslant 34)}{\text{accident\_hors\_voiture ET } (\text{âge} \leqslant 34)}$$
$$= \frac{113 + 75 + 129 + 83}{(541 - 163) + (595 - 354) + (2822 - 2112) + (2353 - 1376)} \simeq \ldots\ldots\ldots\ldots\ldots 0.173.$$

2. NON, car on ne connait pas la taille de la population française en 2005.

Exercice 3.

1. Les 6 valves fonctionnent de manière indépendante. La i-ème valve peut se trouver dans un des deux états : V_i = la valve est ouverte ou $\bar{V}_i$ = la valve est fermée. La probabilité que les 3 valves, V_1, V_2, V_3 soient ouvertes est donc
$$\mathbb{P}(V_1 \text{ et } V_2 \text{ et } V_3) = \mathbb{P}(V_1)\mathbb{P}(V_2)\mathbb{P}(V_3) = \Big(\frac{9}{10}\Big)^3 = 0.729.$$

2. Une canalisation est ouverte si toutes les valves se trouvant sur son chemin sont ouvertes. Le nombre de canalisations ouvertes est noté X ; il peut prendre comme valeur $X = 0$, $X = 1$, $X = 2$ et $X = 3$.

   (a) $(X = 0)$ est l'intersection de 3 événements indépendants
   $$(\bar{V}_1 \text{ ou } \bar{V}_2 \text{ ou } \bar{V}_3) \text{ et } (\bar{V}_4 \text{ ou } \bar{V}_5) \text{ et}(\bar{V}_6)$$
   c'est-à-dire, désigne le fait que, simultanément, une des valves 1 ou 2 ou 3 est fermée, et que, une des valves 4 ou 5 est fermée, et que la valve 6 est fermée. On peut simplifier le calcul en constatant que l'événement contraire à $(\bar{V}_1$ ou $\bar{V}_2$ ou $\bar{V}_3)$, est $(V_1$ et V_2 et $V_3)$, qu'on note pour simplifier $V_1V_2V_3$. On obtient donc,
   $$\begin{aligned}\mathbb{P}(X = 0) &= \big(1 - \mathbb{P}(V_1V_2V_3)\big)\big(1 - \mathbb{P}(V_4V_5)\big)\big(1 - \mathbb{P}(V_6)\big)\\ &= \Big(1 - \Big(\frac{9}{10}\Big)^3\Big)\Big(1 - \Big(\frac{9}{10}\Big)^2\Big)\Big(1 - \Big(\frac{9}{10}\Big)\Big) \simeq 0.005.\end{aligned}$$

   (b) $(X = 1)$ est la réunion des 3 événements disjoints,
   $$(V_1V_2V_3 \text{ et } \overline{V_4V_5} \text{ et } \bar{V}_6) \text{ ou } (\overline{V_1V_2V_3} \text{ et } V_4V_5 \text{ et } \bar{V}_6) \text{ ou } (\overline{V_1V_2V_3} \text{ et } \overline{V_5V_6} \text{ et } V_6).$$
   On obtient donc
   $$\begin{aligned}\mathbb{P}(V_1V_2V_3 \text{ et } \overline{V_4V_5} \text{ et } \bar{V}_6) &= \Big(\frac{9}{10}\Big)^3\Big(1 - \Big(\frac{9}{10}\Big)^2\Big)\Big(1 - \Big(\frac{9}{10}\Big)\Big),\\ \mathbb{P}(\overline{V_1V_2V_3} \text{ et } V_4V_5 \text{ et } \bar{V}_6) &= \Big(1 - \Big(\frac{9}{10}\Big)^3\Big)\Big(\frac{9}{10}\Big)^2\Big(1 - \Big(\frac{9}{10}\Big)\Big),\\ \mathbb{P}(\overline{V_1V_2V_3} \text{ et } et\overline{V_4V_5} \text{ et } V_6) &= \Big(1 - \Big(\frac{9}{10}\Big)^3\Big)\Big(1 - \Big(\frac{9}{10}\Big)^2\Big)\Big(\frac{9}{10}\Big),\\ \mathbb{P}(X = 1) &= \mathbb{P}(V_1V_2V_3 \text{ et } \overline{V_4V_5} \text{ et } \bar{V}_6) + \mathbb{P}(\overline{V_1V_2V_3} \text{ et } V_4V_5 \text{ et } \bar{V}_6)\\ &\quad + \mathbb{P}(\overline{V_1V_2V_3} \text{ et } et\overline{V_4V_5} \text{ et } V_6),\\ &= 0.013851 + 0.021951 + 0.046341 \simeq 0.082.\end{aligned}$$

   (c) $(X = 2)$ est la réunion des 3 événements disjoints
   $$(V_1V_2V_3 \text{ et } V_4V_5 \text{ et } \bar{V}_6) \text{ ou } (V_1V_2V_3 \text{ et } \overline{V_4V_5} \text{ et } V_6) \text{ ou } (\overline{V_1V_2V_3} \text{ et } V_4V_5 \text{ et } V_6).$$
   On obtient donc
   $$\begin{aligned}\mathbb{P}(X = 2) &= \Big(\frac{9}{10}\Big)^3\Big(\frac{9}{10}\Big)^2\Big(1 - \Big(\frac{9}{10}\Big)\Big) + \Big(\frac{9}{10}\Big)^3\Big(1 - \Big(\frac{9}{10}\Big)^2\Big)\Big(\frac{9}{10}\Big)\\ &\quad + \Big(1 - \Big(\frac{9}{10}\Big)^3\Big)\Big(\frac{9}{10}\Big)^2\Big(\frac{9}{10}\Big)\\ &= 0.059049 + 0.124659 + 0.197559 \simeq 0.381.\end{aligned}$$

   (d) $(X = 3)$ est l'événement $V_1V_2V_3V_4V_5V_6$ de probabilité
   $$\mathbb{P}(X = 3) = \Big(\frac{9}{10}\Big)^6 \simeq 0.531.$$

On réunit l'ensemble de ces informations dans un tableau

$\boldsymbol{k}$	0	1	2	3
$\mathbb{P}(\boldsymbol{X}=\boldsymbol{k})$	0.005	0.082	0.382	0.532

(La dernièrer valeur a été ajustée pour obtenir une somme égale à 1).

3. L'événement « l'eau s'écoule de A vers B » est $(X \geqslant 1)$. Sa probabilité est donc

$$\mathbb{P}(X \geqslant 1) = 1 - \mathbb{P}(X = 0) = 1 - 0.005 = 0.995.$$

Exercice 4. On appelle X une variable aléatoire de loi normale $\mathcal{N}(1.005, 0.010^2)$. Une réalisation de X représente un roulement à bille pris au hasard.

1. Comme toujours on revient à une variable normalisée $\mathcal{N}(0,1)$, qu'on notera Z,

$$\begin{aligned}\mathbb{P}(\text{« le roulement à bille est conforme »}) &= \mathbb{P}(0.98 \leqslant X \leqslant 1.02)\\ &= \mathbb{P}\Big(\frac{0.98-1.005}{0.010} \leqslant \frac{X-1.005}{0.010} \leqslant \frac{1.02-1.005}{0.010}\Big) = \mathbb{P}(-2.5 \leqslant Z \leqslant 1.5).\end{aligned}$$

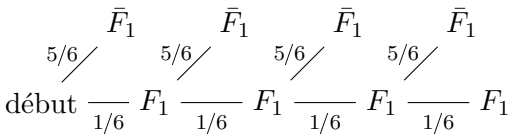
En utilisant les tables de la loi normale, on obtient

$$\begin{aligned}\mathbb{P}(\text{« le roulement à bille est conforme »}) &= \mathbb{P}(Z \leqslant 1.5) - (1 - \mathbb{P}(Z \leqslant 2.5))\\ &= 0.9332 - (1 - 0.9938) = 0.927.\end{aligned}$$

La probabilité que le roulement ne soit pas conforme est donc $1 - 0.927 = 0.073$.

2. Il faudrait que la moyenne μ vérifie $\mu = 1.000$.

Exercice 5. Le jeu peut se résumer par un arbre : la face F_1 arrive avec probabilité $1/6$; une face différente $\bar{F}_1$ arrive avec probabilité $5/6$; les feuilles de l'arbre correspondent aux différents états du jeu en fin de partie.



1. On note X le nombre de fois que la face 1 est tirée. Alors $X \in \{0,1,2,3,4\}$. La loi de X est la donnée des probabilités $\mathbb{P}(X=k)$. Par exemple

$\mathbb{P}(X=0) = \mathbb{P}(\text{« la face 1 ne sort pas au premier lancé »})$,

$\mathbb{P}(X=1) = \mathbb{P}(\text{ « la face 1 sort au premier lancé mais ne sort pas au deuxième lancé »})$.

D'où $\mathbb{P}(X=0) = \frac{5}{6}$ et $\mathbb{P}(X=1) = \frac{1}{6} \times \frac{5}{6}$. On regroupe tous les résultats dans un tableau

$\boldsymbol{X}=\boldsymbol{k}$	0	1	2	3	4
$\mathbb{P}(\boldsymbol{X}=\boldsymbol{k})$	$\frac{5}{6}$	$\frac{5}{6^2}$	$\frac{5}{6^3}$	$\frac{5}{6^4}$	$\frac{1}{6^4}$

2. Le bénéfice Y dépend de la valeur de X. Si par exemple $X=0$, le joueur gagne (ou plutôt perd) $Y(0) = 80 - 100$ €, si $X=1$, il gagne $Y(1) = 3 \times 80 - 100$ €. On regroupe ces calculs dans un tableau

$\boldsymbol{X}=\boldsymbol{k}$	0	1	2	3	4
$\boldsymbol{Y}(\boldsymbol{k})$	$80-100$	$3\times 80-100$	$3^2\times 80-100$	$3^3\times 80-100$	-100
	-20	140	620	2060	-100

L'espérance du gain est donc

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{k=0}^{4} \mathbb{P}(X=k)Y(k)\\ &= \frac{5}{6}\times(-20) + \frac{5}{36}\times 140 + \frac{5}{216}\times 620 + \frac{5}{1296}\times 2060 + \frac{1}{1296}\times(-100) = 25 \text{ €}.\end{aligned}$$

Un joueur qui répéterait ce jeux un grand nombre de fois gagnerait en moyenne 25 € à chaque partie.

D.11 Enoncé de l'examen de juin 2014

Exercice 1. On considère une pièce éclairée par 8 lampes, fonctionnant de manière indépendante, de rendement distribué selon une loi normale de paramètres μ (moyenne) et σ (écart-type) exprimés en lumens par watt. On dira qu'une pièce respecte la réglementation si le rendement moyen $\bar{X}$ des 8 lampes est supérieur à 10 lm/W

$$\bar{X} = \frac{1}{8}\big(X_1 + \cdots + X_8\big), \quad X_i = \text{ rendement de la ième lampe.}$$

On supposera dans les deux questions suivantes $\sigma = 0.7$ lm/W.

1. On suppose que le rendement moyen d'une lampe est de $\mu = 9.8$ lm/W.
   (a) Déterminer la moyenne de $\bar{X}$, $\mathbb{E}[\bar{X}]$, et l'écart-type de $\bar{X}$, $\sqrt{\mathrm{Var}(\bar{X})}$. (On admettra par la suite que $\bar{X}$ suit une loi normale $N(\mu, \sigma/\sqrt{n})$).
   (b) Quelle est la probabilité que la pièce respecte la réglementation ?
2. Quel rendement moyen μ par lampe faut-il choisir pour que la réglementation soit respectée 8 fois sur 10 ?

Exercice 2. Une enquête est réalisée auprès de N mineurs licenciés afin de déterminer lesquels d'entre eux ont pu retrouver un emploi.

1. (Indépendant de la suite.) Dans une enquête réalisée en se déplaçant à domicile, 1 chômeur sur 5 refuse de répondre au questionnaire. 12 visites sont réalisées un certain jour. Déterminer la probabilité que :
   (a) 2 personnes exactement refusent de répondre,
   (b) au moins 9 personnes acceptent de répondre au questionnaire.
2. On préfère réaliser l'enquête par courrier. On note N le nombre de questionnaires envoyés et n le nombre questionnaires retournés.
   (a) On suppose que $N = 85$ et $n = 27$. Déterminer un intervalle de confiance au seuil de 95% de la proportion des mineurs qui répondent au questionnaire.
   (b) On note p la proportion qu'un mineur au chômage, qui a répondu au questionnaire, retrouve un emploi. Sur les 27 questionnaires retournés, 11 d'entre eux ont retrouvé un emploi. Déterminer le nombre de questionnaires n qu'il aurait fallu récupérer à la place de 27 pour que la largeur de l'intervalle de confiance de la proportion p (calculée au seuil de 95%) soit égale à 0.2.

Exercice 3. Un échantillon de 16 vaches laitières Prim'Holstein a été suivi pendant une semaine. La production de lait en kilogrammes est recensée dans le tableau suivant

167.3	144.8	101.9	115.3	122.7	146.4	157.1	127.8
102.6	168.4	115.9	131.1	144.7	169.2	155.4	119.3

Déterminer un intervalle de confiance au seuil de 95% de la production moyenne de lait pendant une semaine des vaches Prim'Holstein. (On donnera les formules théoriques, le nom de la loi utilisée, le détail du calcul numérique, et les deux bornes de l'intervalle de confiance.)

Exercice 4. Un microbiologiste désire savoir s'il existe une différence de temps dans la fabrication de yaourts en utilisant deux types de ferments : le ferment A *lactobacillus acidophilus* ou le ferment B *lactobacillus bulgaricus*. Le tableau suivant mesure en heures le temps de fabrication pour 14 lots : 7 lots utilisant le ferment A et 7 lots utilisant le ferment B,

Ferment A	6.5	6.8	7.2	6.2	8.1	7.0	6.1
Ferment B	6.1	6.2	5.7	6.9	6.3	6.8	5.9

1. Peut-on affirmer que le temps de fabrication en utilisant A est différent de celui en utilisant B, au risque de 5% ?
2. Déterminer la p-valeur de ce test. (Un encadrement suffira.)

Exercice 5. On s'intéresse à la défaillance d'un certain serveur informatique. On comptabilise chaque jour le nombre d'incidents de ce serveur sur une période de 260 jours. Le tableau suivant donne le nombre de fois que k incidents journaliers se soient produits, pour $k = 0, 1, 2, 3$ et $k \geqslant 4$, sur la période de 260 jours,

Nombre d'incidents	0	1	2	3	4 (et plus)	
Nombre de jours	71	88	58	29	14	← **Total :260**

On cherche à savoir si cette distribution peut être modélisée par une distribution de Poisson pour un certain paramètre λ. On rappelle la forme de la distribution d'une variable de Poisson X

Nombre d'incidents k	0	1	2	3	4 (et plus)
Probabilité $\mathbb{P}(\mathbf{X = k})$	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{1}{2}\lambda^2 e^{-\lambda}$	$\frac{1}{6}\lambda^3 e^{-\lambda}$	$1 - \mathbb{P}(X \leqslant 3)$

On rappelle aussi que $\mathbb{E}[X] = \lambda$ pour une telle loi. (Aucune autre connaissance n'est exigée pour manipuler cette loi.)

1. En se servant du premier tableau, déterminer la moyenne observée $\hat{\lambda}$ du nombre d'incidents journaliers. (Pour la dernière classe, on supposera que 4 incidents journaliers et pas plus se sont produits 14 fois sur 260 jours).
2. En se servant du second tableau et du résultat de la question précédente, déterminer ce que serait la distribution du nombre d'incidents journaliers si elle était de Poisson de paramètre $\hat{\lambda}$. (La réponse devra donner un tableau de 3 lignes : première ligne, le nombre d'incidents k, deuxième ligne, $\mathbb{P}(X = k)$, troisième ligne, le nombre de fois théorique que k incidents journaliers se soient produits sur 260 jours.)
3. Au risque de $\alpha = 5\%$, peut-on affirmer que la distribution observée du nombre d'incidents ne suit pas une distribution de Poisson de paramètre $\hat{\lambda}$? (On rédigera très proprement les étapes d'un test d'hypothèse à préciser.)
4. A quel seuil de signification peut-on affirmer que la distribution est de Poisson ? (Un encadrement de la p-valeur suffira.)
5. (Bonus) Expliquer en quelques lignes pourquoi le degré de liberté de ce problème n'est pas $ddl = 4$ comme le suggère le cours, mais $ddl = 3$. (Il ne s'agit pas de reprendre les calculs précédents avec ce nouveau ddl).

D.12 Solution de l'examen de juin 2014

Exercice 1.

1. (a) Indépendamment de la loi de X, on a toujours, $\mathbb{E}[\bar{X}] = \mathbb{E}[X_1]$ et $\mathrm{Var}(\bar{X}) = \frac{1}{8}\mathrm{Var}(X_1)$. On obtient ainsi nuériquement

$$\mu = \mathbb{E}[\bar{X}] = 9.8 \text{ lm/W} \quad \text{et} \quad \sigma_{\bar{X}} = 0.25 \text{ lm/W}.$$

   (b) La probabilité p que la réglementation soit respectée est

$$\begin{aligned} p = \mathbb{P}(\bar{X} \geqslant 10) &= \mathbb{P}\Big(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \geqslant \frac{10 - \mu}{\sigma_{\bar{X}}}\Big) \\ &= \mathbb{P}\Big(\mathcal{N}(0,1) \geqslant \frac{10 - 9.8}{0.25} = 0.81\Big) \simeq 1 - 0.7910 \simeq 0.209. \end{aligned}$$

   (On a utilisé l'égalité en loi $\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \sim \mathcal{N}(0,1)$, ainsi que la table de fonction de répartition de la loi normale). D'où $p \simeq 21\%$.

2. Réciproquement, si on demandait quelle plus petite moyenne μ il faudrait choisir pour avoir $p = 80\%$, on devrait résoudre

$$p = 80\% = \mathbb{P}\Big(\mathcal{N}(0,1) \geqslant \frac{10 - \mu}{\sigma_{\bar{X}}}\Big).$$

On remarque que cela entraîne que le quantile $q = \frac{10-\mu}{\sigma_{\bar{X}}}$ est négatif ; $\mathbb{P}(\mathcal{N} < q) = 20\%$), par symétrie de la loi $\mathbb{P}(\mathcal{N} > -q) = 20\%$, et donc $\mathbb{P}(|\mathcal{N}| > -q) = 40\%$. La table des quantiles de la loi normale donne

$$-q = \frac{\mu - 10}{\sigma_{\bar{X}}} = 0.842, \quad \mu = 10 + 0.842 * 0.25 \simeq 10.21.$$

Exercices 2.

1. On note X le nombre de chômeurs refusant de répondre au questionnaire. La loi de X est donc une loi binomilae $\mathcal{B}(n,p)$ avec $n = 12$ et $p = \frac{1}{5} = 20\%$. La formule générale est $\mathbb{P}(X = k) = \binom{n}{k}(\frac{1}{5})^k(\frac{4}{5})^{n-k}$.
   (a) Deux personnes exactement refuse de répondre

$$\mathbb{P}(X = 2) = \binom{12}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^{10} \simeq 28\%.$$

   (b) Au moins 9 personnes acceptent de répondre, c'est-à-dire, au plus 3 personnes refusent de répondre

$$\begin{aligned}\mathbb{P}(X \leqslant 3) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3),\\ &\simeq 6.9\% + 20.6\% + 28.3\% + 23.6\% \simeq 79\%.\end{aligned}$$

2. (a) La proportion observée p_{obs} de questionnaires retrounés est donnée par $p_{obs} = \frac{n}{N}$; son intervalle de confiance au seuil $1-\alpha$ est donc $p = p_{obs} \pm z_\alpha \frac{\sigma_{obs}}{\sqrt{N}} = p_{obs} \pm z_\alpha \sqrt{\frac{p_{obs}(1-p_{obs})}{N}}$. Numériquement on obtient : $N = 85$, $n = 27$, $p_{obs} \simeq 0.32$, $\alpha = 5\%$, $z_\alpha \simeq 1.96$, $\sigma_{obs} \simeq 0.05$, d'où

$$22\% \lesssim p \lesssim 42\%.$$

   (b) On observe une proportion $p = \frac{11}{27}$ de personnes ayant retrouvé un emploi. On aimerait connaître le nombre de questionnaire n qu'il aurait fallu récupérer pour que la largeur de l'intervalle de confiance soit égale à 0.2. On cherche donc à résoudre l'équation

$$2z_\alpha\sqrt{\frac{p_{obs}(1-p_{obs})}{n}} = 0.2 \quad \Longleftrightarrow \quad n = \left(\frac{2z_\alpha}{0.2}\right)^2 p_{obs}(1-p_{obs}).$$

   Numériquement on obtient $p \simeq 0.41$, et donc $n \simeq 93$.

Exercice 3. Il s'agit d'un intervalle de confiance d'une moyenne. Si μ_{obs} et σ_{obs} désigne la moyenne et l'écart-type observés, l'intervalle est donnée par $\mu = \mu_{obs} \pm t_\alpha \frac{\sigma_{obs}}{\sqrt{n}}$, où t_α désigne le quantile absolu d'une loi de Student pour un degré de liberté $dll = 16 - 1 = 15$. On obtient numériquement, $n = 16$, $\mu_{obs} \simeq 136.9$, $\sigma_{obs} \simeq 22.9$, $\alpha = 5\%$, $t_\alpha \simeq 2.131$. D'où

$$124.6 \leqslant \mu \leqslant 149.1.$$

Exercice 4. On est en présence d'un test de comparaison de moyenne. On note

$H_0 =$ « les temps de fabrication ne diffèrent pas d'un ferment à l'autre ». L'ensemble de rejet est alors

$$\mathcal{R} = \left\{\frac{|\bar{X}_A - \bar{X}_B|}{S_{AB}\sqrt{1/n_A + 1/n_B}} > t_\alpha\right\},$$

où $\bar{X}_A$ et $\bar{X}_B$ désignent les estimateurs des moyennes de chaque ferment et S_{AB} l'estimateur de l'écart-type commun aux écarts-types de chaque ferment

$$S_{AB} = \sqrt{\frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{n_A + n_B - 2}}.$$

et t_α, le quantile absolue de la loi de Student au $ddl = n_A + n_B - 2$.

1. Numériquement on obtient, $n_A = n_B = 7$, $\bar{x}_A \simeq 6.84$, $\bar{x}_B \simeq 6.27$, $s_A \simeq 0.67$, $s_B \simeq 0.44$, $s_{AB} \simeq 0.58$, $\alpha = 5\%$, $dll = 12$, $t_\alpha \simeq 2.179$, d'où

$$t_{crit} = \frac{|\bar{x}_A - \bar{x}_B|}{s_{AB}\sqrt{1/n_A + 1/n_B}} \simeq 1.854.$$

Comme $t_{crit} < t_\alpha$, on ne peut pas rejeter H_0 au niveau de confiance 95% : les moyennes observées de temps de fabrication ne sont pas significativement différentes.

2. Pour pouvoir rejeter H_0 au vu des résultats, il faudrait choisir $\alpha = p_{valeur}$ de sorte que $t_{crit} = t_\alpha$. La table des quantiles absolues de la loi de Student donne l'encadrement

$$1.782 \leqslant t_{crit} \leqslant 2.179, \quad \text{d'où} \quad 5\% \leqslant \alpha \leqslant 10\%.$$

(Un calcul plus précis utilisant un logiciel de calcul donnerait $p_{valeur} \simeq 8.9\%$: au risque de se tromper 9 fosi sur 100, on peut affirmer que les temps moyens de fermentation sont différents).

Exercice 5.

1. Si N_k désigne le nombre de fois, observé sur une période de $N = 260$ jours, que k incidents journaliers se produisent, la moyenne observée du nombre d'incidents journaliers est alors

$$\hat{\lambda} = \sum_{k=0}^{4} k\frac{N_k}{N} = 0 * \frac{71}{260} + 1 * \frac{88}{260} + 2 * \frac{58}{260} + 3 * \frac{29}{260} + 4 * \frac{14}{260} \simeq 1.335.$$

2. On note k, le nombre d'incidents, $p_k = \mathbb{P}(X = k)$, la probabilité que k incidents se produisent si la loi était de Poisson, et $p_k N$, le nombre de fois théorique que k incidents journaliers arrivent (toujours sous l'hypothèse d'une loi de Poisson). On réunit l'ensemble de ces informations dans le tableau suivant

$\boldsymbol{k}$	0	1	2	3	4
$\boldsymbol{p_k}$	0.263	0.351	0.234	0.104	0.047
$\boldsymbol{p_k N}$	68.45	91.35	60.96	27.12	12.12

3. On cherche à savoir si la distribution observée est différente de la distribution de Poisson de paramètre $\hat{\lambda}$ trouvé dans la première question. Il s'agit d'un test d'hypothèse d'ajustement. On note

$$H_0 = \text{« la distribution du nombre d'incidents journaliers est de Poisson de paramètre } \hat{\lambda} \text{ »}.$$

L'ensemble de rejet est $\mathcal{R} = \{D > q_{1-\alpha}\}$, où $D = \sum_{k=0}^{4}(N_k - p_k N)^2/p_k N$ et $q_{1-\alpha}$ est le quantile $1-\alpha$ à $dll = 5 - 1 = 4$ de la loi du chi-deux. Numériquement, on obtient pour $\alpha = 5\%$,

$$q_{95\%} = 9.488 \quad \text{et} \quad D = 0.783.$$

En conclusion, $D < q_{95\%}$, on ne peut pas rejeter H_0, ou bien, si on rejetait H_0, on se tromperait plus que 5 fois sur 100. La distribution observée suit donc, apparemment, une loi de Poisson.

4. La p_{valeur} permet d'estimer l'erreur statistique qu'on commet en disant que les deux distributions diffèrent. La table des quantiles du chi-deux à 4 degrés de liberté donne

$$0.711 \leqslant D \leqslant 1.064 \quad \text{d'où} \quad 90\% \leqslant p_{valeur} \leqslant 95\%.$$

(Un calcul sur logiciel donnerait $p_{valeur} = 94\%$). En disant que la distribution observée n'est pas de Poisson, on se trompe 94 fois sur 100 : c'est une manière forte d'affirmer, cette fois-ci, que les deux distributions coïncident.

5. (Hors programme). Dans la définition de l'estimateur,

$$D = \sum_{k=0}^{4} \frac{(N_k - p_k N)^2}{p_k N}$$

les variables $Z_k = N_k - p_k N$ ne sont pas indépendantes mais reliées entre elles par deux relations linéaires (au lieu d'une seule dans le cas du cours)

$$\sum_{k=0}^{4} Z_k = 0 \quad (\text{correspondant à } N = \sum_{k=0}^{4} N_k), \quad \text{et} \quad \sum_{k=0}^{4} k Z_k = 0,$$

correspondant au fait que la moyenne $\hat{\lambda}$ est aussi un estimateur (au lieu d'être une constante),

$$\hat{\lambda} = \sum_{k=0}^{4} \frac{N_k}{N} = \sum_{k=0}^{4} k p_k.$$

Le degré de liberté est donc 5-2=3 au lieu de 5-1=4.

E Solutions des exercices corrigés

E.1 Espace et mesure de probabilité

Solution 1 (Exercice 17)**.** (a) Ω est l'ensemble des arrangements avec répétition de 3 réels pris dans $\{1,2,3,4,5,6\}$. Son cardinal est égal à $6^3 = 216$.

(b) Il y a 6 brelans possibles d'où $\mathbb{P}(\text{brelan}) = 1/36$.

(c) Une paire est une issue de la forme AAX, AXA, XAA, en désignant par A, X deux éléments distincts de $\{1,2,3,4,5,6\}$. Ces trois famille de paires ont même cardinal. Il y a $6 \times 5 = 30$ choix pour le couple (A, X), il y a donc $3 \times 30 = 90$ paires dans Ω. D'où $\mathbb{P}(\text{paire}) = 90/216 = 5/12$.

(d) L'événement "trois faces distinctes" est le complémentaire de "un brelan ou une paire". D'où

$$\mathbb{P}(\text{trois faces distinctes}) = 1 - \mathbb{P}(\text{un brelan}) - \mathbb{P}(\text{une paire}) = \frac{5}{9}.$$

Solution 2 (Exercice 17)**.** Les hypothèses de l'énoncé nous permettent de remplir le diagramme 13.

(1) Il y a donc exactement 6 touristes sur 100 choisissant exclusivement S1 et S2.

(2) Il y a 5 chance sur 100 qu'un touriste ne choisisse aucune visite.

(3) La probabilité que deux visites exclusivement soient choisies est égale à $5\% + 15\% + 6\% = 26\%$.

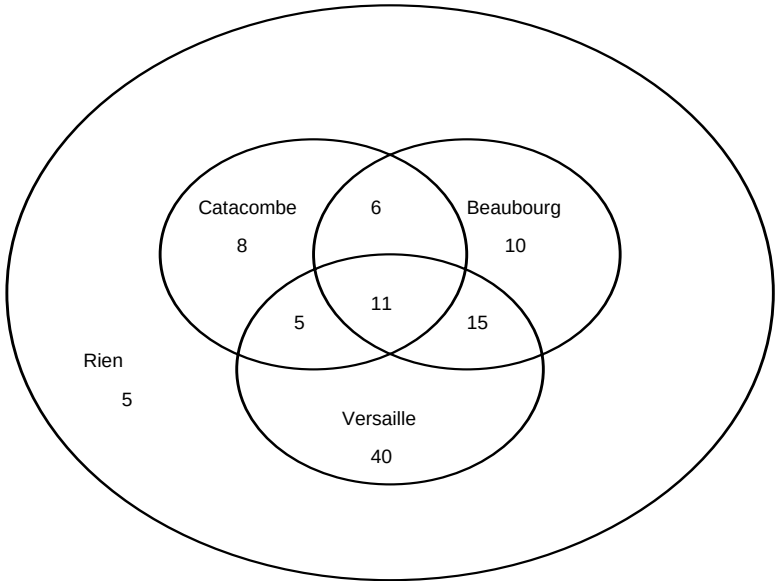


FIGURE 13 – Diagramme de répartition

E.2 Indépendance et probabilités conditionnelles

Solution 3 (Exercice 29)**.** On cherche à établir des statistiques sur un carrefour très dangereux. On observe pour cela un grand nombre de conducteurs empruntant ce carrefour. Un événement élémentaire est décrit par deux informations : le conducteur a (ou n'a pas) eu un accident à ce carrefour, le conducteur a (ou n'a pas) moins de 25 ans. L'ensemble fondamental est donc égal à

$$\Omega = \{AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}\},$$

où par exemple $\bar{A}B$ désigne l'ensemble des conducteurs de moins de 25 ans n'ayant pas eu d'accident.

(1) On donne par hypothèse $\mathbb{P}(B) = \frac{1}{5} = 20\%$, $\mathbb{P}(\bar{B}) = \frac{4}{5} = 80\%$, $\mathbb{P}(A|B) = 3\%$ et $\mathbb{P}(A|\bar{B}) = 2\%$. On peut alors remplir le tableau suivant :

	B	$\bar{B}$
A	0.6%	$1,6\%$
$\bar{A}$	19.4%	78.4%

D'où $\mathbb{P}(A) = \mathbb{P}(\text{«avoir un accident »}) = (0.6 + 1.6)\% = 2.2\%$.

(2) Si on sait par avance qu'un accident a eu lieu, la probabilité qu'il ait été commis par un conducteur agé de moins de 25 ans est $\mathbb{P}(B|A) = \mathbb{P}(AB)/\mathbb{P}(A) = 0.6/2.2 = 27\% > 20\% = \mathbb{P}(B)$.

Solution 4 (Exercice 30)**.** A chaque CD-Rom on peut associer deux informations : le CD-Rom est défectueux ou non, le CD-Rom est éliminé ou non. L'ensemble fondamental est donc constitué de tous les CD-Rom qu'on peut ranger de deux manières différentes : dans D ou son complémentaire $\bar{D}$, dans E ou son complémentaire $\bar{E}$. D'où

$$\Omega = \{DE,\ D\bar{E},\ \bar{D}E,\ \bar{D}\bar{E}\}.$$

Les CD-Rom défectueux, qu'ils soient éliminés ou non, sont rassemblés dans $D = \{DE,\ D\bar{E}\}$. L'énoncé donne les informations suivantes :

$$\mathbb{P}(D) = 30\%, \quad \mathbb{P}(E|D) = 90\%, \quad \mathbb{P}(\bar{E}|\bar{D}) = 80\%.$$

(1)

$$\begin{aligned}
\mathbb{P}(DE) &= \mathbb{P}(E|D)\mathbb{P}(D) = 90\% * 30\% = 27\%,\\
\mathbb{P}(\bar{D}\bar{E}) &= \mathbb{P}(\bar{E}|\bar{D})\mathbb{P}(\bar{D}) = 80\% * 70\% = 56\%,\\
\mathbb{P}(D\bar{E}) &= \mathbb{P}(D) - \mathbb{P}(DE) = 30\% - 27\% = 3\%.
\end{aligned}$$

On peut résumer ces calculs dans un tableau des événements simultanés :

Ω	E	$\bar{E}$	
D	27%	3%	30%
$\bar{D}$	14%	56%	70%
	41%	49%	100%

(2) La probabilité qu'un CD-Rom soit élimié est donné par

$$\mathbb{P}(E) = 41\%.$$

(3) Si un CD-Rom est éliminé, la probabilité qu'il soit déffectueux est donné par

$$\mathbb{P}(D|E) = \frac{\mathbb{P}(DE)}{\mathbb{P}(E)} = \frac{27}{41} = 66\%.$$

Solution 5 (Exercice 31)**.** La motion est acceptée s'il y a au moins 3 personnes parmi les 6 membres à avoir voté « oui ». Si à chacun des membres, $i = 1, 2 \cdots, 6$, on associe une v.a. de Bernoulli X_i, valant 1 lorsque la personne vote « oui » et 0 sinon, le nombre total de personnes membres votant « oui » est donc égal à $X = X_1 + X_2 + \cdots + X_6$. La variable X suit une loi de binomiale de paramètre $n = 6$, $p = 0.4$. On demande de calculer $\mathbb{P}(X \geqslant 3)$, on a :

$$\begin{cases} \mathbb{P}(X = 0) & = (0.6)^6 = 4.7 \\ \mathbb{P}(X = 1) & = 6 * (0.4) * (0.6)^5 = 18.7 \\ \mathbb{P}(X = 2) & = 15 * (0.4)^2 * (0.6)^4 = 31.1 \end{cases} \qquad \begin{cases} \mathbb{P}(X \leqslant 2) & = 54.4 \\ \mathbb{P}(X \geqslant 3) & = 45.6 \end{cases}$$

Solution 6 (Exercice 32)**.** (1) Sans aucune autre information supplémentaire, un lot a une chance sur douze d'être produit un mois donné. d'où

$$\mathbb{P}(\text{"le lot provient des 3 premiers mois"}) = 3/12 = 0.25.$$

(2) Chaque montre peut être considérée comme une v.a. à laquelle on associe deux informations : (1) elle appartient à un lot produit au cours des 3 premiers mois, ou son contraire ; (2) elle est défectueuse, ou son contraire. L'énoncé permet de construire un arbre de probabilités en commençant par calculer la probabilité des deux éventualités de l'information (1) puis celle des deux éventualités de l'information (2) sachant celle de (1). On obtient ainsi le tableau :

	non défectueuse	**défectueuse**
lot des 3 premiers mois	$(3/12) * 60\%$	$(3/12) * 40\%$
lot des 9 derniers mois	$(9/12) * 90\%$	$(9/12) * 10\%$

Ce tableau permet de calculer d'abord la probabilité qu'une montre soit défectueuse, quelle provienne des 3 premiers mois ou des 9 derniers :

$$\mathbb{P}(\text{“défectueuse”}) = (3/12) * 40\% + (9/12) * 10\% = 21/120.$$

Il permet ensuite de calculer la probabilité conditionnelle

$$\mathbb{P}(\text{“lot des 3 premiers mois”} \mid \text{“défectueuse”}) = \frac{(3/12) * 40\%}{21/120} = \frac{12}{21} = 0.57.$$

E.3 Variables aléatoires discrètes et lois usuelles

Solution 7 (Exercice 47)**.** (i) Soit X le nombre de garçons parmi 4 personnes. Les valeurs possibles de X sont $X = 0, 1, 2, 3, 4$. Si $X = 0$, les 4 personnes sont toutes des filles ; si $X = 1$, : il y a $\binom{4}{1}$ possibilités de choisir 1 garçon parmi 4 personnes ; si $X = 2$, il y a $\binom{4}{2}$ possibilités de choisir 2 garçons parmi 4 personnes, ainsi de suite. Comme la probabilité de choisir une fille ou une garçon est de $\frac{1}{2}$, X suit donc la loi binomiale$B(n, p)$ avec $n = 4$ et $p = \frac{1}{2}$, soit $\mathbb{P}(X = k) = \binom{4}{k}(\frac{1}{2})^4$. On obtient ainsi le tableau

k	0	1	2	3	4
$\mathbb{P}(X = k)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

(ii) On commnece par calculer, pour chaque valeur de X, le nombre Y de couples possibles. Par exemple, si $X = 1$, il y a 1 garçon et 3 filles, on peut donc former 1 couple (il y a bien 3 manières de former ce couple). On obtient le tableau

X	0	1	2	3	4
Y	0	1	2	1	0

D'où la loi de Y

l	0	1	2
$\mathbb{P}(Y = l)$	$\frac{2}{16}$	$\frac{8}{16}$	$\frac{6}{16}$

(iii) L'espérance de Y est donnée par la formule

$$\mathbb{E}(Y) = 0 * \mathbb{P}(Y = 0) + 1 * \mathbb{P}(Y = 1) + 2 * \mathbb{P}(Y = 2).$$

On obtient ici $\mathbb{E}(Y) = \frac{20}{16} = \frac{5}{4} = 1,25$. De même

$$\mathbb{E}(Y^2) = 0^2 * \mathbb{P}(Y = 0) + 1^2 * \mathbb{P}(Y = 1) + 2^2 * \mathbb{P}(Y = 2).$$

On a $\mathbb{E}(Y^2) = \frac{32}{16} = 2$. La variance de Y est donc donnée par

$$\text{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = 2 - \frac{25}{16} = \frac{7}{16}.$$

Solution 8 (Exercice 48)**.** En tirant au hasard 7 personnes, ces personnes deviennent des v.a. indépendantes pouvant prendre 3 valeurs : "asiatique", "blanc" et "noir" avec les probabilités 60%, 39% et 1%.

(1) Si on ne s'intéresse qu'au seul fait d'être ou de pas être asiatique, les 7 personnes précédentes peuvent être considérées comme des variables de Bernoulli prenant les 2 valeurs 1 ou 0, soit "asiatique" ou "non asiatique", avec les probabilités 60% et 40%. La somme S de ces v.a. suit une loi binomiale $N(n = 7, p = 60\%)$. La probabilité $\mathbb{P}$ qu'il y ait une majorité d'asiatique est donc égale à

$$\begin{aligned}\mathbb{P} &= \mathbb{P}(S = 4) + \mathbb{P}(S = 5) + \mathbb{P}(S = 6) + \mathbb{P}(S = 7)\\ &= \binom{7}{4}(60\%)^4(40\%)^3 + \binom{7}{5}(60\%)^5(40\%)^2\\ &+ \binom{7}{6}(60\%)^6(40\%) + \binom{7}{7}(60\%)^7\\ &= 71\%.\end{aligned}$$

On a utilisé $\binom{7}{4} = 35$, $\binom{7}{5} = 21$ et $\binom{7}{1} = 7$.

(2) De même, si on ne s'intéresse qu'au seul fait d'être "noir" ou "non noir" avec les probabilités 1% ou 99%, la probabilité $\mathbb{P}$ de n'obtenir aucun noir dans l'échantillon est donc

$$\mathbb{P} = (99\%)^7 = 93\%.$$

Solution 9 (Exercice 49)**.** On appellera par la suite R, D et B les variables aléatoires "recette", "dépense" et "bénéfice". Ces variables dépendent en effet du carnet de commande de l'entreprise et sont donc bien des v.a.

(1) Dans les cas où il n'est pas demandé d'heures supplémentaires aux ouvriers, les dépenses moyennes de l'entreprise sont constantes :

$$\mathbb{E}(D) = D = 5 * 40 * 20 = 4000 \quad \text{euros.}$$

Les recettes sont par contre variables et se calculent à partir du tableau de statistique : on constatera que l'entreprise est limitée à $200 = 5 * 40$ heures de travail, peu importe l'offre du marché. On trouve

$$\mathbb{E}(R) = (3\% * 180 + 9\% * 190 + 88\% * 200) * 30 = 5955 \quad \text{euros.}$$

D'où un bénéfice moyen : $\mathbb{E}(B) = 5955 - 4000 = 1955$ euros.

(2) On recommence le calcule précédent avec 6 ouvriers travaillant chacun 40 heures sans heures supplémentaires. Les dépenses sont encore constantes égales à

$$\mathbb{E}(D) = D = 6 * 40 * 20 = 4800 \quad \text{euros.}$$

L'entreprise est ici limitée à $6 * 40 = 240$ heures de travail. La moyenne des recettes est alors donnée par

$$\begin{aligned}\mathbb{E}(R) = (&3\% * 180 + 9\% * 190 + 12\% * 200 + 15\% * 210 + 22\% * 220\\ &+ 21\% * 230 + 18\% * 240) * 30 = 6537 \quad \text{euros.}\end{aligned}$$

D'où un bénéfice moyen : $\mathbb{E}(B) = 6537 - 4800 = 1737$ euros. Cette opération est moins rentable que dans le premier cas.

(3) Ici les dépenses ne sont pas constantes et dépendent du nombre d'heures supplémentaires. Le volume horaire maximum que l'entreprise peut accepter est de $200 + 4 * 5 = 220$ heures. On a pour les dépenses :

$$\mathbb{E}(D) = 200 * 20 + (15\% * 10 + 61\% * 20) * 25 = 4342 \quad \text{euros.}$$

Pour les recettes :

$$\begin{aligned}\mathbb{E}(R) = (&3\% * 180 + 9\% * 190 + 12\% * 200 \\ &+ 15\% * 210 + 61\% * 220) * 30 = 6366 \quad \text{euros.}\end{aligned}$$

D'où un bénéfice $\mathbb{E}(B) = 6366 - 4342 = 2023$ euros. Il est légèrement supérieure au premier cas.

(4) En remplaçant 4 heures par 6 heures, l'entreprise peut espérer 230 heures de travail. Pour les dépenses et recettes, on a

$$\begin{aligned}\mathbb{E}(D) &= 200 * 20 \\ &+ (15\% * 10 + 22\% * 20 + 39\% * 30) * 25 = 4440 \quad \text{euros.} \\ \mathbb{E}(R) &= (3\%180 + 9\% * 190 + 12\% * 200 \\ &+ 15\% * 210 + 22\%220 + 39\%230) * 30 = 6483 \quad \text{euros.}\end{aligned}$$

D'où un bénéfice $\mathbb{E}(B) = 6483 - 4440 = 2043$ euros sensiblement égale au bénéfice de la question (3).

E.4 Variables aléatoires continues et lois usuelles

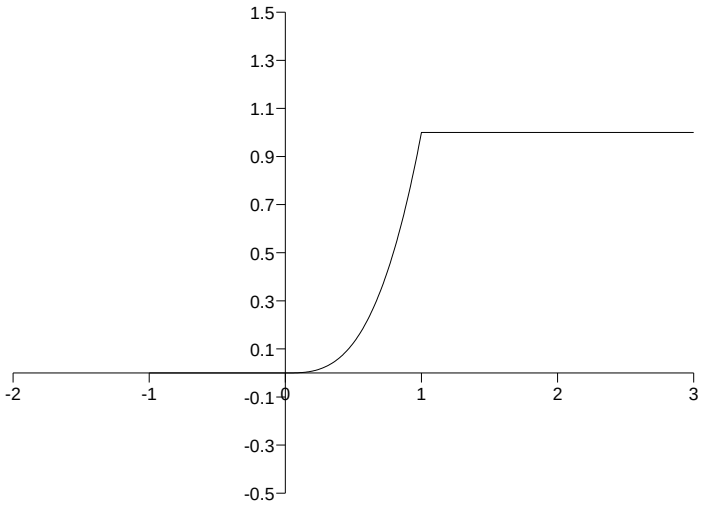
Solution 10 (Exercice 59)**.** (1) Pour que $f(x)$ soit une densité de probabilité, il est nécessaire (et suffisant) que $f(x) \geqslant 0$ partout et $\int f(x)\,dx = 1$. Pour que cette deuxième propriété soit réalisée, il faut

$$\int_0^1 kx^2\,dx = 1 \Leftrightarrow k = 3.$$

(2) Par définition de la fonction de répartition $F(x)$ (ou distribution cumulée) $F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\,du$. Comme $f(x) = 0$ pour $x \leqslant 0$ et $x \geqslant 1$, le calcul de $F(x)$ dépendra de l'intervalle où x se trouve :

- si $x \leqslant 0$, $F(x) = 0$,
- si $0 \leqslant x \leqslant 1$, $F(x) = \int_0^x f(u)\,du = x^3$,
- si $x \geqslant 1$, $F(x) = 1$.

Le graphe de $F(x)$ est donc



(3) Par définition de l'espérance $\mathbb{E}(X) = \int_0^1 3x^3\,dx = \frac{3}{4} = 0.75$. Pour le calcul de la variance, on commence par $\mathbb{E}(X^2) = \int_0^1 3x^4\,dx = \frac{3}{5} = 0.6$. Enfin $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{3}{5} - (\frac{3}{4})^2 = \frac{3}{80} = 0.0375$, soit un écart-type de $\sigma = 0.1936$.

Solution 11 (Exercice 60)**.** (1) On appelle X la tension (exprimée en Volts) de l'alimentation. On suppose par hypothèse que c'est une v.a. de loi normale $N(m, \alpha^2)$ avec $m = 24$ et $\alpha = 1.8$.

(a) Le composant fonctionne normalement si $22 < X < 26$. Sa probabilité p est donc (Z désigne une v.a. normale centrée réduite)

$$\begin{aligned} p &:= \mathbb{P}(\text{«le composant fonctionne»}) \\ &= \mathbb{P}(22 < X < 26) = \mathbb{P}(\frac{22-m}{\alpha} < \frac{X-m}{\alpha} < \frac{26-m}{\alpha}) \\ &= \mathbb{P}(-\frac{2}{1.8} < Z < \frac{2}{1.8}) = \mathbb{P}(-1.11 < Z < 1.11), \end{aligned}$$

Les tables donnent la fonction de répartition de Z, soit $F(x)$. On commence alors par transformer l'expression :

$$\begin{aligned} \mathbb{P}(-x < Z < x) &= 1 - \mathbb{P}(|Z| > x) = 1 - 2\mathbb{P}(Z > x) \\ &= 1 - 2(1 - F(x)) = 2F(x) - 1. \end{aligned}$$

Comme $F(1.11) = 0.8665$, $p = \mathbb{P}(-1.11 < Z < 1.11) = 73\%$.

(b) Le composant est détruit si $X > 29$. Sa probabilité q vaut donc

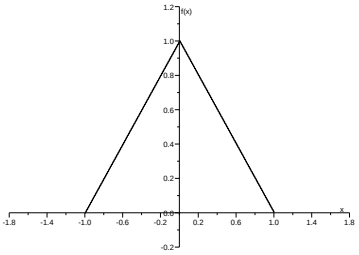
$$\begin{aligned} q &:= \mathbb{P}(\text{«le composant est détruit»}) \\ &= \mathbb{P}(X > 29) = \mathbb{P}(\frac{X-m}{\alpha} > \frac{29-m}{\alpha}) \\ &= \mathbb{P}(Z > \frac{5}{1.8}) = \mathbb{P}(Z > 2.78) = 1 - F(2,78) \\ &= 1 - 0.9973 = 2.7/1000. \end{aligned}$$

(2) On applique un raisonnement à l'envers. On cherche α pour que le composant fonctionne avec une probabilité de $p = 85\%$ (la valeur moyenne m est inchangée) :

$$p = 85\% = \mathbb{P}(\text{«le composant fonctionne»}) = \mathbb{P}(-\frac{2}{\alpha} < Z < \frac{2}{\alpha}).$$

De manière équivalent $\mathbb{P}(|Z| > \frac{2}{\alpha}) = 15\%$. En utilisant les tables, on obtient donc $\frac{2}{\alpha} = 1.44$, soit $\alpha = 1.39$. L'écart-type est bien plus petit que le précédent puisque le composant a plus de chance de fonctionner.

Solution 12 (Exercice 61)**.** (i) Le graphe de f est



(ii) Pour que $f(x)$ soit une densité d'une variable aléatoire, il faut que $\int_{-1}^{1} f(x)\,dx = 1$, soit $b = 1$.

(iii) L'espérance d'une variable Z est donnée par

$$\mathbb{E}(Z) = \int_{-1}^{1} xf(x)\,dx = \int_{-1}^{0} x(1+x)\,dx + \int_{0}^{1} x(1-x)\,dx = 0.$$

(iv) On calcule d'abord

$$\mathbb{E}(Z^2) = \int_{-1}^{1} x^2 f(x)\,dx = \int_{-1}^{0} x^2(1+x)\,dx + \int_{0}^{1} x^2(1-x)\,dx = \frac{1}{6}.$$

Puis on obtient $\mathrm{Var}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2 = \frac{1}{6}$, $\sigma(Z) = \frac{1}{\sqrt{6}}$.

Solution 13 (Exercice 62)**.** (1) Sur chacun des intervalles $]-\infty,-3[$, $[-3,0]$, $[0,3]$ et $[3,\infty[$, le graphe de $f(x)$ est une droite. Il suffit donc de tracer les points $(x, y = f(x))$ en chaque extrémité de ces intervalles. On trouve

La fonction $f(x)$ est la densité d'une v.a. si $f(x) \geqslant 0$ et si $\int f(x)\,dx = 1$. On a

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-3}^{3} f(x)\,dx = 9k = 1 \quad \Longleftrightarrow \quad k = \frac{1}{9}.$$

Pour le calcul de l'intégrale, on remarque que $\int f(x)\,dx$ représente l'aire du domaine situé sous la courbe. C'est l'aire d'un triangle qu'on peut calculer en utilisant la formule classique $aire = \frac{1}{2}\ base \times hauteur$.

(2) Par définition, la fonction de répartition est donnée par

$$F(x) = \int_{-\infty}^{x} f(u)\,du.$$

Il faut distinger 4 cas :

— si $x < -3$, $f(x) = 0$ pour $u \in]-\infty, x]$ et donc $F(x) = 0$,

— si $x > 3$, $F(x) = \int_{-3}^{3} f(u)\,du = 1$,

— si $x \in [-3,0]$, $F(x) = \int_{-3}^{x} f(u)\,du$

$$F(x) = \frac{1}{2}\ base \times hauteur = \frac{1}{2}(x+3)f(x) = \frac{1}{18}(x+3)^2,$$

— si $x \in [0,3]$, $F(x) = \int_{-3}^{0} f(u)\,du + \int_{0}^{x} f(u)\,du$

$$\begin{aligned}\int_0^x f(u)\,du &= \frac{1}{2}\ base \times (bord\ gauche + bord\ droit)\\ &= \frac{1}{2}x\left(\frac{1}{3} + \frac{3-x}{9}\right) = \frac{1}{18}x(6-x)\\ F(x) &= \frac{1}{2} + \frac{1}{18}x(6-x).\end{aligned}$$

Le graphe a donc l'allure suivante (il est composé de deux morceaux de paraboles)

(3) On demande de calculer $\mathbb{P}(X > 2)$

$$\mathbb{P}(X > 2) = \frac{1}{2}(3-2)f(2) = \frac{1}{18} = 5.6\%.$$

Solution 14 (Exercice 63)**.** (1) On suppose que l'automobiliste choisit la deuxième option. Pour chaque jour de travail on définit une variable de Bernoulli X_i valant 1 si l'automobiliste est verbalisé, et 0 sinon. Par hypothèse, la probabilité d'être verbalisé est de 20%. Le nombre total de fois que l'automobiliste est verbalisé est donc égal à

$$X = X_1 + X_2 + \cdots + X_{225}.$$

X suit une loi binomiale $B(n,p)$ de paramètre $n = 225$ et $p = 0.2$. On sait alors calculer

$$\mathbb{E}(X) = np = 45, \quad \mathrm{Var}(X) = np(1-p) = 36.$$

(2) Le total des amendes est égal à $Y = 10X$. Alors

$$\mu = \mathbb{E}(Y) = 450, \quad \mathrm{Var}(Y) = 3600, \quad \sigma = 60.$$

(3) On suppose maintenant que Y suit la loi normale $N(\mu, \sigma)$. On introduit la variable centrée réduite correspondante $Z = (Y - \mu)/\sigma$. Alors

$$\begin{aligned}\mathbb{P}(450 < Y < 458) &= \mathbb{P}(0 < Z < \frac{458 - 450}{60}) = \mathbb{P}(0 < Z < \frac{8}{60}) \\ &= F(0.13) - F(0) = 0.5517 - 0.5 \simeq 5\%\end{aligned}$$

(4) On suppose ici que $S = 500$. L'automobiliste est gagnant si $\mathbb{P}(Y < S) \geqslant 75\%$:

$$\begin{aligned}\mathbb{P}(Y < S) &= \mathbb{P}(Z < \frac{S - \mu}{\sigma}) = \mathbb{P}(Z < \frac{50}{60}) \\ &= \mathbb{P}(Z < 0.83) = 0.7967 \simeq 80\%.\end{aligned}$$

L'automobiliste est donc gagnant.

(5) On cherche maintenant S pour que $\mathbb{P}(Y > S) \geqslant 75\%$. On se ramène encore à la variable centrée réduite Z :

$$25\% = \mathbb{P}(Y < S) = \mathbb{P}(Z < \frac{S - \mu}{\sigma}) = \mathbb{P}(Z < z_0)$$

où on a posé $z_0 = (S - \mu)/\sigma$. On remarque d'abord que z_0 est nécessairement négatif ($25\% < 50\%$). La table ne donne cependant que la probabilité de dépassement de la variable symétrique $|Z|$. On transforme donc le calcul précédent

$$\mathbb{P}(Z > -z_0) = 0.25, \quad \mathbb{P}(|Z| > -z_0) = 0.5, \quad z_0 = -0.674.$$

D'où une offre promotionnelle $S = 450 - 0.674 * 60 > 409$.

E.5 Théorème de la limite centrale

Solution 15 (Exercice 71)**.** (1) A chaque page d'un livre, on associe une variable de Bernoulli valant 1 si elle est erronée et 0 sinon. Ces variables de Bernoulli peuvent être supposées indépendantes. On note X la variable somme des pages erronées. Alors X suit une loi binomiale $B(n, p)$ de paramètre $n = 300$ et $p = 5\%$. En particulier

$$\mathbb{P}(X = k) = \binom{300}{k} \left(\frac{5}{100}\right)^k \left(\frac{95}{100}\right)^{n-k}.$$

(2) Par définition d'une loi binomiale $\mathbb{E}(X) = np = 300 * (5/100) = 15$, $\mathrm{Var}(X) = np(1 - p) = 300 * (5/100) * (95/100) = 14.25$ et $\sigma_X = 3.77$.

(3) Par le théorème limite centrale, la v.a. X peut être approchée par une loi normale $N(\mu, \sigma)$ de paramètre $\mu = np$ et $\sigma = \sqrt{np(1 - p)}$. La probabilité q de rejeter un livre est donc (on appelle Z une v.a. de loi normale $N(0, 1)$)

$$\begin{aligned}q &:= \mathbb{P}(\text{«le livre est rejeté»}) \\ &= \mathbb{P}(X > 20) = \mathbb{P}(\frac{X - \mu}{\sigma} > \frac{20 - \mu}{\sigma}) \\ &= \mathbb{P}(Z > \frac{20 - 15}{3.77}) = \mathbb{P}(Z > 1.32) \\ &= 1 - F(1.32) = 1 - 0.9066 = 9.3\%.\end{aligned}$$

L'éditeur rejette entre 9 et 10 livres tous les 100 livres.

Solution 16 (Exercice 72)**.** (i) Comme X et Y suivent des lois normales et que ces deux variables sont indépendantes, $W = X + Y$ suit aussi une loi normale de paramètres $\mu_W = \mu_X + \mu_Y$ et $\sigma_W^2 = \sigma_X^2 + \sigma_Y^2$. On a

$$\begin{aligned}\mu_W = \mathbb{E}(W) = \mathbb{E}(X) + \mathbb{E}(Y) = 80 + 10\alpha \text{ kg} \\ \sigma_W = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{144 + 25\alpha^2}.\end{aligned}$$

Pour $\alpha = 1$, $\mu_W = 90$ et $\sigma_W = \sqrt{169} = 13$.

(ii) Comme W_{50} est la somme de 50 variables indépendantes qui suivent la même loi normale $\mathcal{N}(\mu_W, \sigma_W^2)$, W_{50} suit la loi normale $\mathcal{N}(50\mu_W, 50\sigma_W^2)$,

$$\mathbb{E}(W_{50}) = 500(8+\alpha), \quad \sigma(W_{50}) = \sqrt{50}\sqrt{144+25\alpha^2}.$$

(iii) La cabine est en surcharge si $W_{50} \geqslant 4700$. Sa probabilité est donnée par (on a supposé $\alpha = 1$) :

$$\begin{aligned}\mathbb{P}(W_{50} \geqslant 4700) &= \mathbb{P}\Big(N = \frac{W_{50}-\mu(W_{50})}{\sigma(W_{50})} \geqslant \frac{4700-500(8+\alpha)}{\sqrt{50}\sqrt{144+25\alpha^2}}\Big)\\ &= \mathbb{P}(N \geqslant 2.1757) = 1-\mathbb{P}(N<2.1757) = 1-0.985\\ &= 1,5\%.\end{aligned}$$

(iv) On veut maintenant que la probabilité de dépasser 4700 kg soit au plus 1% c'est-à-dire

$$\mathbb{P}(W_{50} \geqslant 4700) = \mathbb{P}(N \geqslant z(\alpha)) = 1\% \text{ où } z(\alpha) = \frac{4700-500(8+\alpha)}{\sqrt{50}\sqrt{144+25\alpha^2}}.$$

Nécessairement, $z(\alpha) = 2,326$ et α est solution de l'équation

$$24,3237\alpha^2 - 70\alpha + 45,1046 = 0 \quad \text{et} \quad 4700-500(8+\alpha) \geqslant 0$$

soit $\alpha = (70-\sqrt{4900-4*24.3237*45.1046})/(2*24.3237) = 0,97$.

Solution 17 (Exercice 73)**.** (1) Appelons X le salaire d'un représentant de commerce. Par hypothèse X suit une loi normale $N(\mu = 400, \sigma = 50)$. En appelant N une loi normale standard, on obtient :

$$\begin{aligned}\mathbb{P}(X>422) &= \mathbb{P}(\frac{X-400}{50} > \frac{422-400}{50})\\ &= \mathbb{P}(N>0.44) = 1-\mathbb{P}(N<0.44) = 1-0.67 = 33\%.\end{aligned}$$

(2) Le fait d'être ou de ne pas être une bonne journée est une variable de Bernoulli. Chaque journée est indépendante des autres. La v.a. Y est donc la somme de 5 v.a. de Bernoulli indépendantes ; Y suit une loi binomiale $B(n=5, p=33\%)$. D'où

$$\begin{aligned}\mathbb{P}(X \geqslant 3) &= \mathbb{P}(X=3)+\mathbb{P}(X=4)+\mathbb{P}(X=5)\\ &= 10(33\%)^3(67\%)^2 + 5(33\%)^4(67\%) + (33\%)^5 = 20\%.\end{aligned}$$

(3) Par hypothèse W est la somme de 275 v.a. de Bernoulli indépendantes.

(a) W suit une loi binomiale $B(n=275, p=33\%)$. Son espérance est donc égale à

$$\mu = \mathbb{E}(W) = np = 275(33\%) = 90 \quad \text{bonnes journées.}$$

Son écart-type est donnée par

$$\begin{aligned}\sigma &= \mathrm{Var}(X)^{1/2} = (np(1-p))^{1/2}\\ &= (275(33\%)(67\%))^{1/2} = 8 \quad \text{bonnes journées.}\end{aligned}$$

(b) Le théorème central limite permet d'approcher $(W-\mu)/\sigma$ par la loi normale $N(\mu, \sigma)$.

(c) En appelant N une v.a. de loi normale standard, on obtient

$$\begin{aligned}\mathbb{P}(75<W<100) &= \mathbb{P}(\frac{75-90}{8} < \frac{W-\mu}{\sigma} < \frac{100-90}{8})\\ &= \mathbb{P}(-1.87<N<1.25)\\ &= \mathbb{P}(N<1.25) - \big(1-\mathbb{P}(N<1.87)\big)\\ &= 0.8944-(1-0.9693) = 86\%.\end{aligned}$$

Solution 18 (Exercice 85)**.**

Solution 19 (Exercice 86)**.**

Solution 20 (Exercice 103)**.**

Solution 21 (Exercice 104)**.**

Index

Références

Ouvrages de base

[1] F. Dress, Probabilités et Statistique, pour les sciences de la vie, Dunod (2002).

[2] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, John Wiley & Sons (2011).

[3] T.H. Wonnacott, R. J. Wonnacott, Statistique, Economica (1995).

Ouvrages plus techniques

[4] A. Agresti, An Introduction to Categorical Data Analysis, Wiley-Interscience, John Wiley & Sons (2007).

[5] D. Foata, A. Fuchs, Calcul des probabilités, Cours, exercices et problèmes corrigés, Dunod (1998).

[6] S.M. Ross, Introduction to Probability Models, Academic Press, Elsevier (2007).

[7] G. Saporta. Probabilités, Analyse des données et Statistique. Editions Technip, Paris, 1990.