Fiche 1 - TISD - Master Pro
Rappels de probabilité

Tran Viet Chi, chi.tran@math.univ-lille1.fr, bureau 316 (bâtiment M3).

1 Une introduction au logiciel R

1.1 Séquences, Vecteurs, Tableaux (arrays)

1.1.1 Déclaration et définition

1. Saisir le vecteur a en tapant `a <- c(10, 5, 3, 6, 21)`. Taper a. Puis $a[2]$ et $a[1,3]$.

2. Saisir le vecteur b en tapant cette fois `b<- array(data=c(15, 3, 12, 2, 1),dim=c(1,5))`. Demander b.

3. Pour voir la différence entre une liste et un vecteur, demander `nrow(a)`, `ncol(a)`, `dim(a)` et de même pour b.

4. Générer un vecteur c de dimension 5 dont toutes les composantes sont 1 en utilisant la commande `array`.

5. Générer un vecteur d en tapant `d<-seq(from=1, to=10, by=2)`.

6. Si on veut générer des séquences d'entiers consécutifs, on peut utiliser la commande `1:5`. Créer un vecteur e de dimension 5 et de composantes la séquence des 5 premiers entiers.

7. Générer une matrice diagonale de dimension 3 dont les éléments de la diagonale sont 1, 5 et 9 à l'aide de la commande `diag`.

1.1.2 Manipulations de base

1. Taper `2*a+b+1`. Qu'obtient-on ?

2. Taper `e[3]`. Qu'obtient-on ?

3. Taper `cos(a)`, `exp(a)`. Qu'obtient-on ?

4. Taper `a*e` puis `a%*%e`. Comparer.

5. Taper maintenant `b%*%e`. Que se passe-t-il ? Corriger par `f<- t(b)%*%e`.

6. Demander `dim(f)`. Taper `f[2,3]`, `f[,3]`, `f[2:5,]` puis `f[2:3,4]`. Qu'obtient-on ?

7. Taper `cbind(b,e)` puis `rbind(b,e)`. Pourquoi faut-il éviter d'utiliser `cbind` ou `rbind` avec des objets du type de a ? (Essayer `cbind(a,b)` et `rbind(a,b)`).

8. Taper `c(a,b)`. Qu'obtient-on ?

1.2 Listes

1. Taper

```
Lst <- list(name="Fred", wife="Mary", no.children=3,child.ages=c(9,7,4)).
```

Demander `Lst`.

2. Demander `Lst[[1]]`. Retrouver le résultat en tapant `Lst$name`

3. Demander `Lst[[4]]`. Qu'obtient-on. Demander l'âge du troisième enfant.

1.2.1 Exercices

Exercice 1 (Résolution d'un système linéaire et inversion d'une matrice carrée)
Soient :

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 3 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

1. Saisir la matrice A (par colonnes) et le vecteur b. La matrice A est-elle inversible ? Calculer son déterminant grace à la commande `det(A)`.

2. Résoudre $Ax = b$ avec la commande `solve(A,b)`.

3. Pour `eigen(A, symmetric=FALSE, only.values = FALSE)`. Le résultat de cette commande est une liste. En extraire les valeurs propres, dans une matrice que l'on appellera D et les vecteurs propres dans une matrice de passage que l'on appellera P.

4. calculer P^{-1} à l'aide de la commande `solve(P)` et retrouver A à partir de P, D et P^{-1}.

2 Rappels de probabilité

2.1 Combinatoire avec R

Exercice 2 (Attribution de tableaux)
Si 10 tableaux noirs doivent être affectés à 4 écoles, de combien de manières peut-on les répartir ? Qu'en est-il si chaque école doit recevoir au moins un tableau ? Faire les applications numériques en utilisant la commande `factorial`, puis en utilisant la commande `choose`.

Exercice 3 (Question existentielle concernant un forfait téléphone)
Béatrice a souscrit une formule "appels illimités vers trois numéros de téléphone". Elle doit choisir parmi ses six amis Bertrand, Jean, Marc, Marie, Ouassila et Radu. Enumérer toutes les possibilités en utilisant la commande `combn`. Combien y a-t-il de combinaisons ?

2.2 Quelques lois de probabilité usuelles : lois discrètes

Exercice 4 (Loi de Bernoulli, lois binômiales)

Partie A Introduction

1. Rappeler la définition de la loi de Bernoulli $\mathcal{B}(1,p)$ et des lois binômiales $\mathcal{B}(n,p)$.

2. On cherche à représenter sur un même graphique les probabilités $\mathbb{P}(X = k)$ en fonction de $k \in \mathbb{N}$ pour les lois $\mathcal{B}(j, 0.4)$, j variant de 1 à 5.
a. Expliquer pourquoi il suffit de se restreindre à $k \in [\![0,5]\!]$. Définir un vecteur x dont les composantes sont les entiers de 0 à 5.
b. En utilisant des boucles du type `for(j in 1:6){...}` et des conditions du type `if(j<=i+1){...}` définir une matrice y de dimension 5×6 telle que $\forall i \in [\![1,5]\!]$, $\forall j \in [\![1,6]\!]$, $y_{i,j} = \mathbb{P}(X_i = j-1)$ avec $X_i \rightsquigarrow \mathcal{B}(i, 0.4)$.
c. Faire la représentation graphique en utilisant les fonctions `plot` et `points` avec l'option `type="h"`(on utilisera des couleurs, et afin de ne pas superposer les barres, on pourra utiliser comme abscisses x, $x + 0.03$).

3. Représenter sur un même graphique les fonctions de répartition des lois $\mathcal{B}(1, 0.4)$, $\mathcal{B}(5, 0.4)$ en utilisant les fonctions `plot` et `lines` avec l'option `type="s"`. On pourra utiliser la commande `sum`. Pourquoi ce choix de fonction en escalier ? (modifier l'axe des ordonnées afin que l'intervalle $[0, 1]$ soit représenté).

Partie B Election présidentielle

On suppose qu'une proportion $p \in [0,1]$ de la population compte voter pour Nicolas tandis que les $1-p$ restants ont l'intention de voter pour Ségolène. On interroge $n = 1000$ personnes, choisies de façon indépendante dans la population, et on suppose qu'elles répondent honnêtement. A chaque répondant $i \in [\![1, n]\!]$, on associe une variable aléatoire X_i qui vaut 1 s'il compte voter pour Nicolas et 0 s'il compte voter pour Ségolène. Ces variables aléatoires sont donc supposées *iid* de loi $\mathcal{B}(1,p)$.

1. (théorique) On considère la moyenne empirique $\bar{X}_n = \sum_{i=1}^n X_i/n$. Calculer son espérance, sa variance et donner sa limite lorsque $n \to +\infty$. $\bar{X}_n$ est une approximation de p.

2. Quelle est la loi de $\sum_{i=1}^n X_i$? Dessiner l'histogramme de $N = 3000$ simulations de variables *iid* de même loi que $\sum_{i=1}^n X_i$ pour $p = 0.5$. Dessiner en fonction de $p \in [0,1]$ la probabilité pour que $\bar{X}_n > 1/2$. Commenter.

3. Nous nous intéressons à la probabilité $\mathbb{P}(\bar{X}_n \leq p - 0.01)$. Est-il possible de calculer explicitement cette probabilité ? Numériquement, dessiner cette probabilité en fonction de $p \in [0,1]$. Que vaut-elle pour $p = 1\%$, 50%, 75% ? Quelles sont les valeurs maximales et minimales ?

4. Pour $a > 0$, en utilisant l'inégalité de Bienaymé-Tchebychev, montrez que :

$$\mathbb{P}\left(\left|\bar{X}_n - p\right| \geq a\right) \leq \frac{1}{4a^2 n}$$

4.1. Quel est nombre d'individus n' à interroger pour que la probabilité que l'écart entre $\bar{X}_n$ et p soit supérieur à $a = 1\%$ soit inférieure à 5% ?
4.2. En utilisant le théorème central limite, donner un intervalle fonction de $\bar{X}_n$ contenant p avec probabilité 0.95 lorsque l'on interroge $n = 1000$ personnes. Dans le cas où $\bar{X}_n = 51\%$ (et $1 - \bar{X}_n = 49\%$), pouvez-vous faire un commentaire de ce résultat ?

Cet exercice est inspiré d'un sondage réalisé par la TNS-Sofres le 24 avril 2007. Le cadre de cet exercice est bien sûr simplificateur, par les hypothèses faites sur les répondants et par les techniques d'estimation de p choisies (nécessité de redressement des fausses ou non-réponses, méthodes de sondage par "quota" pour garantir la représentativité de l'échantillon de personnes interrogées...)

Partie C Simulation de fractales

1. Définir les matrices suivantes :

$$A_0 = \begin{pmatrix} 0.839 & -0.303 \\ 0.383 & 0.924 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -0.161 & -0.136 \\ 0.138 & -0.182 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 0.232 \\ -0.080 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0.921 \\ 0.178 \end{pmatrix}$$

Puis simuler la suite suivante :

$$\forall n \geq 1,\ X_{n+1} = A(n)\, X_n + B(n), \quad X_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{1}$$

où $(A(n),\, B(n)) = (A_0,\, B_0)$avec probabilité 0.9 et $(A(n),\, B(n)) = (A_1,\, B_1)$avec probabilité 0.1 (2)

Utiliser dans le programme N pour le nombre de simulation. On commencera par $N = 1000$ simulations, puis on pourra augmenter N. Après la simulation $i \in [\![1, N]\!]$, tracer les points déjà simulés :

```
plot(t(simul[,1:i]), type='p', pch=21, xlim=c(-0.1, 1), ylim=c(-0.1, 1.1))
points(simul[1,i+1],simul[2,i+1], pch=19, col='red')
```

2. Définir les matrices suivantes :

$$A_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0.16 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0.2 & -0.26 \\ 0.23 & 0.22 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 1.6 \end{pmatrix} \tag{3}$$

$$A_3 = \begin{pmatrix} -0.15 & 0.28 \\ 0.26 & 0.24 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0.85 & 0.04 \\ -0.04 & 0.85 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 \\ 0.44 \end{pmatrix}, \quad B_4 = \begin{pmatrix} 0 \\ 1.6 \end{pmatrix} \tag{4}$$

Puis simuler la suite suivante :

$$\forall n \geq 1,\ X_{n+1} = A(n)\, X_n + B(n), \quad X_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{5}$$

où $(A(n),\, B(n)) = (A_1,\, B_1)$avec probabilité 0.1, $(A_2,\, B_2)$avec proba 0.15, $(A_3,\, B_3)$avec proba 0.15, (6)

$(A_4,\, B_4)$avec proba 0.6. (7)

Tracer comme précédemment les points au fur et à mesure de la simulation :

```
plot(t(simul[,1:i]), type='p', pch='.', xlim=c(-5, 5), ylim=c(-1, 10))
points(simul[1,i+1],simul[2,i+1], pch=19, col='red')
```

Exercice 5 (Lois de Poisson, géométriques et binômiales négatives)

Partie A Loi de Poisson

La loi de Poisson permet par exemple de modéliser la loi du nombre d'occurrence de phénomènes répétitifs sur des intervalles de temps donnés, ces éléments étant séparés par des durées exponentielles *iid* (ex : nombre de pannes dans un système, nombre de passage de trains à une station de métro...)

1. (théorique) Rappeler la définition de la loi de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$. Calculer son espérance et sa variance.

2. Représenter $\mathbb{P}(X = k)$ en fonction de $k \in \mathbb{N}$ pour $\lambda = 1/2$ et $\lambda = 2$.

3. Superposition Soient X et Y deux variables aléatoires indépendantes de lois respectives $\mathcal{P}(\lambda_1)$ et $\mathcal{P}(\lambda_2)$. Quelle est la loi de $X + Y$? Le vérifier numériquement de la façon suivante :

- simuler deux variables aléatoires de Poisson indépendantes et de paramètres $\lambda = 500$ et $\lambda = 300$ et les additionner.
- répéter $n = 1000$ fois cette simulation,
- tracer un *QQ-plot* des simulations obtenues en rapport avec la loi de Poisson adéquate.

Partie B Lois géométriques

Pierre lance un dé truqué qui donne `6` avec probabilité $p \in [0, 1]$. Il effectue des lancers jusqu'à obtenir un `6` et s'arrête alors.

1. (théorique) Quelle est la loi du nombre de lancer ? Ecrire sa définition. Calculer son espérance et sa variance.

2. Représenter $\mathbb{P}(X = k)$ en fonction de $k \in \mathbb{N}^*$ pour $p = 1/2$ et $p = 0.3$.

Partie C Lois binômiales négatives

Armand lance le dé de Pierre et compte le nombre de fois où une face autre que `6` apparaît avant la $r^{\text{ième}}$ occurence d'un `6`. Ce nombre suit une loi binômiale négative $\text{NegBin}(r, p)$.

1. (théorique) Donner la définition de la loi binômiale négative $\text{NegBin}(r, p)$. Calculer son espérance et sa variance. Quelle est cette loi lorsque $r = 1$? Soit $X \rightsquigarrow \text{NegBin}(r, r/(\lambda + r))$; montrer que lorsque $r \to +\infty$, $\mathbb{P}(X = k) \to \lambda^k/(k!)e^{-\lambda}$ pour $k \in \mathbb{N}^*$. Conclusion ?

2. Représenter $\mathbb{P}(X = k)$ en fonction de $k \in \mathbb{N}^*$ pour $p = 0.3$ et $r = 4$. Quel est son mode ? Faire varier p et r.

Exercice 6 (Lois multinômiales et hypergéométriques)

1. Virginie souhaite interroger $n = 100$ personnes avec remise prises parmi $N = 1000$ personnes dont $N_1 = 480$ votent à droite, $N_2 = 450$ votent à gauche, et $N_3 = 80$ s'abstiennent. Soient n_1, n_2 et n_3 le nombre de personnes de chaque groupe (comptées éventuellement avec leurs répétitions) se trouvant dans l'échantillon constitué. Quelle est sa loi ?

2. Simuler une variable aléatoire multinômiale de même loi que (n_1, n_2, n_3) à l'aide de tirages aléatoires uniformes sur $[0, 1]$.

3. Même question s'il s'agit d'un tirage sans remise et si Virginie ne veut constituer un échantillon que de personnes votant à droite ou à gauche. Simuler une variable aléatoire hypergéométrique de même loi que (n_1, n_2).

2.3 Quelques lois usuelles : lois continues

Exercice 7 (Lois uniforme et Beta)

Partie A Lois Beta

Les densités des lois Beta, à support sur $[0,1]$, sont caractérisées par deux paramètres de forme $a>0$ et $b>0$:

$$f(x\,;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\quad \text{où } \Gamma(k)=\int_0^{+\infty}t^{k-1}e^{-t}dt \tag{8}$$

Dessiner ces densités pour :

- $\alpha<1$ et $\beta<1$: forme en "U",
- $\alpha<1$, $\beta\geq 1$ ou $\alpha=1$, $\beta>1$: décroissante, en particulier, pour $\alpha=1$, regarder les différentes formes lorsque $1<\beta<2$, $\beta=2$ ou $\beta>2$.
- $\alpha=1$, $\beta=1$: uniforme,
- $\alpha=1$, $\beta<1$ ou $\alpha>1$, $\beta\leq 1$: croissante, en particulier, pour $\beta=1$, regarder les différentes formes suivant la position de α par rapport à 2.
- $\alpha>1$ et $\beta>1$: unimodale.

Partie B Lois uniformes

Nous souhaitons faire quelques tests simples pour tester le le générateur uniforme de **R**. Des tests plus précis seront considérés à la fin du cours.

1. Pour tester le générateur uniforme de **R**, simuler $n=10,\,100,\,10\,000$ variables aléatoires indépendantes dans la loi $\mathcal{U}([0,1])$ et tracer leur histogramme.

2. Pour $n=10\,000$, compter le nombre n_0 de valeurs dans l'intervalle $[0,0.1[$, le nombre n_1 de valeurs dans l'intervalle $[0.1,0.2[$, ..., le nombre n_9 de valeurs dans l'intervalle $[0.9,1]$. Faire un graphique.

3. On considère l'échantillon $X_1,\cdots X_n$ pour $n=10\,000$ de variables *iid* dans la loi $\mathcal{U}([0,1])$. Tracer les $(n-1)$ points de coordonnées (X_i,X_{i+1}) pour $i\in[\![1,n-1]\!]$. Calculer la corrélation de la série $(X_1,\cdots X_{n-1})$ avec la série $(X_2,\cdots X_n)$.

Exercice 8 (Loi normale)
Les lois normales apparaissent dans de très nombreuses situations pratiques (mesures avec erreurs) et dans le théorème central limite.
1. Rappeler la densité de la loi normale $\mathcal{N}(m,\sigma^2)$ d'espérance m et de variance σ^2. Rappeler l'énoncé du théorème central limite pour une suite de variables *iid*.

2. Dessiner sur un même graphique ces densités en faisant varier l'espérance et la variance : $(m,\sigma^2)\in\{(0,1),(1,1),(0,16)\}$.

3. Donner un intervalle symétrique par rapport à l'origine contenant $\alpha=95\%$ de la masse de la loi $\mathcal{N}(0,1)$. Même question pour $\alpha=50\%$ et $\alpha=99\%$.

4. Simuler $n=1000$ variables aléatoires *iid* de loi $\mathcal{N}(m=5,\sigma^2=100)$. Centrer et réduire les observations obtenues. Faire un *QQ-plot* avec la loi $\mathcal{N}(0,1)$. Calculer la moyenne et l'écart-type des observations centrées réduites.

Exercice 9 (Lois exponentielle et Gamma)

Partie A Lois exponentielles

1. Rappeler la définition de la loi de exponentielle de paramètre $\lambda>0$. Représenter sur un même graphique les densités de cette loi pour $\lambda=0.5$ et $\lambda=2$.

2. Calculer l'expression de la fonction de répartition d'une $\mathcal{E}(\lambda)$ et représenter sur un même graphique ces fonctions de répartition pour $\lambda=0.5$ et $\lambda=2$.

3. Calculer l'expression de la fonction quantile d'une $\mathcal{E}(\lambda)$ et sur le graphique précédent, représenter la fonction quantile pour $\lambda=0.5$ ainsi que la première bissectrice.

4. Propriété de "sans-mémoire" Si $X\rightsquigarrow\mathcal{E}(\lambda)$ et si $0<s<t$, montrer que $\mathbb{P}\left(X>t+s\,|\,X>s\right)=\mathbb{P}\left(X>t\right)$. Vérifier-le sur n simulations *iid* dans la loi $\mathcal{E}(\lambda=2)$, en remplaçant les probabilités par leur équivalent empirique :

$$\frac{\text{card}\{X_i>t+s\}}{\text{card}\{X_i>s\}}\quad\text{et}\quad\frac{\text{card}\{X_i>t\}}{n}$$

pour $t=0.4$ et $s=1$. Tracer ces quantités en fonction de n.

Partie B Lois Gamma

1. Quelle est la loi de la somme de deux variables *iid* $\mathcal{E}(\lambda)$?

2. La loi $\Gamma(k, \theta)$ est caractérisée par un paramètre de forme $k > 0$ et un paramètre d'échelle $\theta > 0$:

$$f(x\,;\,k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} \mathbf{1}_{\mathbb{R}_+^*}(x). \tag{9}$$

Représenter ces densités pour $(k, \theta) \in \{(0.5, 1), (1, 1), (2, 1), (2, 2)\}$.

3. Réaliser la simulation suivante $N = 1000$ fois :
- simuler 3 variables aléatoires *iid* $\mathcal{E}(\lambda = 2)$.
- les additionner, et stocker le résultat Y_i $(i \in [\![1, N]\!])$.

Nous disposons à la fin de cette itération de $N = 1000$ variables aléatoires qui chacune est la somme de trois variables exponentielles indépendantes de même paramètre. Faire un *QQ-plot* pour comparer la distribution empirique de ces variables aléatoires avec une loi Gamma bien choisie.

Exercice 10 (Simulation de variables aléatoires de loi de Weibull avec un générateur uniforme)
La loi de Weibull est utilisée pour modéliser des variables aléatoires positives (par exemple des durées), et son comportement ressemble à celui de la loi normale (mais le support reste $\mathbb{R}_+$).

1. La densité de la loi de Weibull est caractérisée par deux paramètres $a > 0$ (paramètre de forme) et $b > 0$ (paramètre d'échelle) :

$$f(x\,;\,a,b) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left(-\left(\frac{x}{b}\right)^a\right). \tag{10}$$

Dessiner sur un même graphique la densité de la loi de Weibull pour $(a, b) \in \{(0.5, 1), (1, 1), (1, 2), (2, 1)\}$.

2. Etablir l'expression de la fonction de répartition F, puis dessiner pour les paramètres précédents les différentes fonctions de répartition sur un même graphique.

3. (théorique) Calculer les fonctions de hasard et commenter en fonction des valeurs de a.

4. Etablir l'expression de la fonction quantile, puis dessiner sur le même graphique que les fonctions de répartition la fonction quantile pour $(a, b) = (1, 2)$. Ajouter la première bissectrice.

5. Soit U une variable aléatoire de loi $\mathcal{U}[0, 1]$. Quelle est la loi de $F^{-1}(U)$? En déduire une façon de simuler une suite de variables *iid* de loi $\mathcal{W}(a, b)$.