Fiche 2 - TISD - Master Pro
Lois de probabilités

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

1 Combinatoire avec R

Exercice 1 (Attribution de tableaux)
Si 10 tableaux noirs (les tableaux sont identiques) doivent être affectés à 4 écoles, de combien de manières peut-on les répartir si chaque école doit recevoir au moins un tableau ? **Indication:** : on note n_i le nombre de tableaux dans l'école i. Les nombres n_i sont complètement déterminés par les quantités n_1, n_1+n_2, $n_1+n_2+n_3$. Faire les applications numériques en utilisant la commande `factorial`, puis en utilisant la commande `choose`. Maintenant, une école peut ne recevoir aucun tableau. De combien de manières différentes peut-on répartir les tableaux ? **Indication:** : bien prendre en compte les zéros potentiels.

Exercice 2 (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 Quelques lois de probabilité usuelles

R connaît la plupart des lois de probabilités usuelles. On y accède en chargeand le package stats (`library(stats)`). Il peut :
- Simuler une variable suivant cette loi grâce à la commande `rloi`
- Donner la densité : `dloi`
- Donner la fonction de répartition : `ploi`
- Donner la fonction quantile : `qloi`.

Lois de probabilités usuelles connues dans R :
- lois binomiales : binom
- lois multinomiales : multinom
- lois de Poisson : pois
- lois uniformes : unif
- lois normales : norm
- lois exponentielles : exp

Toutes les commandes ci-dessus sont à retenir. R connaît encore beaucoup d'autres lois.

3 Lois discrètes

3.1 Lois binômiales

Exercice 3 (Introduction)

1. Rappeler la définition de la loi de Bernoulli $\mathcal{B}(1,p)$ et des lois binômiales $\mathcal{B}(n,p)$. On cherche à représenter sur un même graphique les probabilités $\mathbb{P}(X=k)$ en fonction de $k \in \mathbb{N}$ pour les lois $\mathcal{B}(i,0.4)$, i variant de 1 à 5.
2. Donner les probabilités $P_{i,k}\mathbb{P}(X=k)$ où $X \sim \mathcal{B}(i,0.4)$.
3. Expliquer pourquoi il suffit de se restreindre à $k \in [\![0,5]\!]$.
4. Utiliser R pour construire cette matrice P de taille 5*6. Plutôt que d'utiliser la formule donnée plus haut, on pourra utiliser les fonctions statistiques de R.
5. Visualiser ces probabilités :

```
plot(x,Y[1,], type="h", col="red")
for(i in 2:5)
{
points(x+i*0.03,Y[i,], type="h")
}
```

   La commande `plot` permet de tracer un premier graphique. La commande `points` permet de supperposer d'autres graphiques (si on réutilisait `plot`, on effacerait le premier graphique. On a décalé les barres pour qu'elles ne se chevauchent pas.
6. 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 les options `type=s""`, `xlim=c(-0.1,5.1)` et `ylim=c(-0.1,1.1)`. On pourra utiliser la commande `cumsum`. Pourquoi ce choix de fonction en escalier ? (modifier l'axe des ordonnées afin que l'intervalle $[0,1]$ soit représenté).

Exercice 4 (Sondage)
On suppose qu'une proportion $p \in [0,1]$ de la population compte voter pour François tandis que les $1-p$ restants ont l'intention de voter pour Martine. On interroge $n = 1000$ personnes, choisies de façon indépendante dans la population (avec remise), 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 François et 0 s'il compte voter pour Martine. 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 $p = 0.5$. Utiliser la commande `rbinom` pour les simulations. On trace un histogramme à l'aide de la fonction `hist`.
3. Qui va gagner ? Donner la probabilité que $\bar{X}_n > 1/2$ en fonction de p.
4. Dessiner en fonction de $p \in [0,1]$ la probabilité pour que $\bar{X}_n > 1/2$ (utiliser les fonctions de la toolbox `statistics`. Commenter.

5. On s'intéresse maintenant à la précision de nos résultats. Nous nous intéressons à la probabilité $\mathbb{P}(\bar{X}_n \notin [p-0.01, p+0.01])$. Donner l'expression exacte de cette probabilité, puis une expression approchée en fonction de la loi normale. Pour quelle(s) valeur(s) de p cette valeur est-elle la plus précise ? La moins précise ?
6. Construire une fonction `test` qui calcule la probabilité exacte en fonction de $p \in [0,1]$ et de n. Construire `test2` qui calcule la probabilité approchée. Si $n = 1000$, que valent ces probabilités pour $p = 1\%$, 50%, 75% ? Conclusion. Même question pour $n = 10000$.

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

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

8. Quel est nombre d'individus $n1$ à interroger pour que la probabilité que l'écart entre $\bar{X}_n$ et p soit supérieur à $a = 1\%$ soit inférieure à 5% si on utilise l'inégalité de Bienaymé-Tchebitchev.
9. Quel est nombre d'individus $n2$ à interroger pour que la probabilité que l'écart entre $\bar{X}_n$ et p soit supérieur à $a = 1\%$ soit inférieure à 5% si on utilise le théorème centrale limite.
10. 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. Pour l'application numérique, on pourra utiliser la fonction `qnorm` de R. 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...)

Exercice 5 (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 en créant une fonction dans R :

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

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

Utiliser dans le programme N pour le nombre de simulation. On commencera par $N = 10$ simulations, puis on pourra augmenter N. Après la simulation $i \in [\![1, N]\!]$, tracer les points déjà simulés :

```
plot(X[1,1],X[2,1], type='p', xlim=c(-0.1, 1), ylim=c(-0.1, 1.1))
points(X[1,k],X[2,k])
```

Quand on augment N, on peut faire une pause dans le programme (il suffira d'appuyer sur la touche entrée) grâce à la commande à `scan()`. N'utilisez pas cette commande à chaque itération !

3.2 Lois de Poisson

Exercice 6 ()
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.

4 Lois continues

4.1 Lois uniformes

Exercice 7 ()
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 $(X_2, \cdots X_n)$.

4.2 Loi normale

Exercice 8 ()
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*. Que se passe-t-il dans le cas de lois normales iid ?
2. Dessiner sur un même graphique ces densités en faisant varier l'espérance et la variance : $(m, \sigma^2) \in \{(0,1), (1,1), (0,4)\}$.
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 $\mathcal{N}(0,1)$. Superposer la première bissectrice. Calculer la moyenne et l'écart-type des observations centrées réduites.

4.3 Loi exponentielle

Exercice 9 ()

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 $\mathbb{P}(X > t + s \,|\, X > s) = \mathbb{P}(X > t)$.
5. 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

4.3.1 Lois gamma

Exercice 10 ()

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

$$\mathbb{P}(X + Y \in dx) = \int_0^x \mathbb{P}(X \in du)\mathbb{P}(Y \in d(x - u))du$$

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

$$f(x\,;\, k, \theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} \mathbf{1}_{\mathbb{R}_+^*}(x) \tag{3}$$

Où, pour un entier, $\Gamma(k) = (k - 1)!$.

2. Montrer qu'une loi exponentielle de paramètre λ est une loi Gamma. Préciser les paramètres de cette loi.
3. Montrer que la somme de deux exponentielles de paramètre λ est une loi Gamma. Préciser les paramètres.
4. Représenter la densité d'une loi Gamma pour $(k, \theta) \in \{(0.5, 1), (1, 1), (2, 1), (2, 2)\}$.

5. Réaliser la simulation suivante $N = 1000$ fois :
   - simuler 3 variables aléatoires *iid* $\mathcal{E}(\lambda = 2)$.
   - les additionner, et stocker le résultat Y_i ($i \in [\![1, N]\!]$).
6. Quelle est la loi de la somme des trois variables exponentielles ?
   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.
7. Faire un *QQ-plot* pour comparer la distribution empirique de ces variables aléatoires avec une loi Gamma bien choisie. Ne pas oublier de superposer la première bissectrice.

5 Autres lois discrètes

5.1 Lois géométriques et binômiales négatives

Exercice 11 (Avec un dé)

Partie A 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 B 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 12 (Sondage)

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. 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). Remarque : lorsqu'on fait un sondage sans remise dans une petite population, on ne peut plus supposer que les données sont indépendantes.

5.2 Lois continues

Exercice 13 (Lois Beta)

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

$$f(x\,;\,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad \text{où } \Gamma(k) = \int_0^{+\infty} t^{k-1}e^{-t}dt \tag{4}$$

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.