Université Paris Descartes
UFR de Mathématiques et Informatique

Introduction aux probabilités - Licence MIA 2e année - Parcours Informatique

Travaux Pratiques - 2ème séance

1 Compléments sur R

1.1 Programmation

Les structures de contrôle sous Pascal (alternatives et répétitives) existent sous "R", avec une syntaxe très proche

- `if(`*`booléen`*`) {...} else {...}`
- `for(`*`variable`* `in` *`vecteur`*`) {...}`
- `while(`*`booléen`*`) {...}`
- `repeat {...}`

Remarques :

1. L'instruction `break` permet de terminer une boucle ; c'est la seule façon de quitter une boucle `repeat`.
2. L'instruction `next` permet de passer directement au cycle suivant.
3. Deux commandes successives doivent être séparées par un point-virgule, ou être sur deux lignes distinctes.
4. Les connecteurs logiques sont `!` pour "non", `&` ou `&&` pour "et", `|` ou `||` pour "ou", et `xor` pour "ou exclusif".
5. Les prédicats élémentaires sont `==`, `!=`, `<=`, `<`, `>=`, et `>`.

1.2 Lois de probabilité usuelles

R est un logiciel de statistiques qui permet de manipuler simplement les lois de probabilité classiques : loi binomiale, loi géométrique, loi normale, etc. Pour chacune de ces lois, il est possible d'obtenir au moyen des préfixes `r`, `d`, `p`, et `q`, respectivement des tirages aléatoires, les probabilités discrètes ou la densité, la fonction de répartition et les quantiles. Par exemple, supposons que X soit une variable de loi de Poisson de paramètre $\lambda = 2.5$. Alors

`> rpois(10,2.5)`

renvoie dix tirages aléatoires suivant la loi de X,

`> dpois(4,2.5)`

renvoie la valeur de $P(X = 4)$, c'est-à-dire $e^{-2.5}\frac{2.5^4}{4!}$,

`> ppois(4,2.5)`

renvoie la valeur de $F_X(4) = P(X \leq 4)$,

`> qpois(0.6,2.5)`

renvoie la valeur $\inf\{t, F_X(t) > 0.6\}$.

Voici quelques exemples de lois prédéfinies en R :

Loi	Appelation	Arguments	Valeurs par défaut
Binomiale de paramètre $(n,q) \in \mathbb{N}^* \times [0,1]$	`binom`	`size=n` `prob=q`	
Binomiale négative de paramètre $(n,q) \in \mathbb{N}^* \times [0,1]$	`nbinom`	`size=n` `prob=q`	
Géométrique de paramètre $q \in [0,1]$	`geom`	`prob=q`	
Normale (ou Gaussienne) de paramètre $(\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+$	`norm`	`mean=`μ `sd=`σ	`mean=0` `sd=1`
Poisson de paramètre $\lambda \in \mathbb{R}_+$	`pois`	`lambda=`λ	
Uniforme sur $[min, max]$	`unif`	`min=`min `max=`max	`min=0` `max=1`

Exercice 1. *Déterminer une fonction* `poisson_intervalle`*, qui à* $\lambda > 0$ *et deux réels* $a < b$*, renvoie* $\mathbb{P}(a < X \leq b)$*, pour* X *suivant une loi de Poisson de paramètre* $\lambda > 0$*. Même question pour* $\mathbb{P}(a \leq X \leq b)$*.*

Exercice 2.

1. *Donner la commande qui, à un paramètre* $q \in]0,1[$*, renvoie une réalisation de la variable aléatoire* X *qui suit une loi géométrique de paramètre* q*.*
2. *A l'aide de variables aléatoires indépendantes de Bernoulli de paramètre* $q \in]0,1[$*, définir une fonction* `geo_2` *qui renvoie une réalisation d'une variable* Y *suivant une loi géométrique de paramètre* q*.*

Exercice 3. *Dans cette exercice on définit deux quantités, la moyenne empirique* $\bar{X}_n$*, et la variance empirique* S_n^2 *respectivement par :*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

où $X_1, \ldots, X_n$ *sont des réalisations i.i.d. d'une variable aléatoire. Dans cette exercice on considèrera des variable de loi exponentielle de paramètre* 2*.*

1. *Rappeler la valeur de* $\mathbb{E}(X_1)$ *et* $\mathbb{V}(X_1)$*.*
2. *Créer une fonction* `resulat` *qui prend en argument* n *et renvoie en sortie les valeurs* $\bar{X}_n$ *et* S_n^2*.*
3. *Calculer pour* $n = 10,\ 100,\ 1000$*, les valeurs* $\bar{X}_n$ *et* S_n^2*. Que remarque-t-on ?*
4. *Comparer les sorties de* `resultat`*, avec ceux des fonctions* `R`*,* `mean` *et* `var`*.*

Exercice 4. *On appelle médiane d'une variable aléatoire X la valeur m définie par* $\inf\{t \in \mathbb{R}, F(X) > \frac{1}{2}\}$ *où* $F(x) = \mathbb{P}(X \leq x)$*, est la fonction de répartition de X. Soient $X_1, \ldots, X_n$ n variables aléatoires i.i.d. de fonction de répartition F. On appelle médiane empirique, $\bar{m}_n$ la valeur*

$$\bar{m}_n = \begin{cases} X_{([\frac{n}{2}])} & \text{si } n \text{ est pair} \\ X_{([\frac{n}{2}]+1)} & \text{si } n \text{ est impair} \end{cases}$$

où $[a]$ désigne la partie entière de a et $X_{(1)} \leq X_{(2)} \leq \ldots X_{(n)}$ est l'échantillon $X_1, \ldots, X_n$ ordonnée par valeurs croissantes.

1. *Que représente la valeur m ?*
2. *Donner la médiane d'une loi uniforme sur $[0, 1]$ à l'aide de la fonction* `qruinf`*.*
3. *Définir une fonction* `mediane_emp` *qui prend en argument n et renvoie la médiane empirique de n variables i.i.d. $(X_1, \ldots, X_n)$ suivant une loi uniforme sur $[0, 1]$.*
4. *Calculer pour $n = 10$, 100, 1000, la valeur $\bar{m}_n$. Que remarque-t-on ?*
5. *Comparer les sorties de* `mediane_emp`*, avec ceux de la fonction* `R`*,* `median`*.*

1.3 Matrices

On peut créer une matrice en R grâce à la commande `matrix` :

```
> M = matrix(data=0,nr=10,nc=5)
```

`M` est alors une matrice de 10 lignes et 5 colonnes contenant des zéros. On peut accéder à ses éléments grâce aux crochets :

```
> M[3,4]
[1] 0
> M[3,4]=7
> M[3,4]
[1] 7
> M
```

On peut sélectionner une sous-matrice en précisant plusieurs indices de lignes et de colonnes :

```
> M[3:6,4:5]=7
> M
```

L'absence d'indice de ligne ou de colonne signifie qu'on les sélectionne toutes :

```
> M[8:10,]=4
> M
```

2 Simulation de file d'attente

Dans cette partie nous allons simuler une file d'attente de clients dans un bureau de poste. La théorie des files d'attentes est un domaine à part entière et vise à fournir des solutions optimales de gestion des files. Les résultats de ce domaine ont trouvé des applications pour les algorithmes d'ordonnancement des systèmes d'exploitations, dans les télécoms pour la gestion des réseaux téléphoniques, ou encore la gestion des départs et arrivées d'avions dans un aéroport.

2.1 Première partie

On suppose qu'à chaque minute un client au maximum peut entrer dans le bureau de poste, avec une probabilité p. Le bureau de poste dispose d'un seul guichet, et on suppose que chaque client passant au guichet y reste pendant un nombre de minutes aléatoire modélisé par une loi géométrique de paramètre q. La poste est ouverte de 9 heure à 17 heure sans interruption.

Nous désignerons par X la variable aléatoire représentant le temps entre l'arrivée de deux clients. Par Y_h nous désignerons la variable aléatoire représentant le nombre de clients entrés en une heure, et par Y_j le nombre de clients entrés en une journée.

Exercice 5. *Quelles lois suivent les variables X, Y_h et Y_j ? Écrivez trois fonctions* `proba_X()`, `proba_Y_h()` *et* `proba_Y_j()` *qui chacune prend en argument une valeur du support de X, Y_h ou Y_j et renvoie la probabilité que la variable prenne cette valeur.*

Nous souhaitons réaliser une simulation d'une journée d'ouverture du bureau de poste.

Exercice 6. *Ecrire une fonction* `sim_day_1` *prenant en argument les probabilités p et q, simule une journée d'ouverture du bureau de poste et qui renvoie le vecteur contenant le nombre de clients présents pour chaque minute entre 9h et 17h.* Conseil : utilisez une boucle `for` itérant autant de fois qu'il y a de minutes de 9 heure à 17 heure.

Exercice 7. *Tester la fonction* `sim_day_1`*, d'abord avec $p = 0.1$ et $q = 0.1$ puis en faisant varier p et q. Dans chaque cas, afficher le graphique du vecteur retourné (fonction* `plot`*) et calculer dans chaque cas la taille maximum et la taille moyenne de la file d'attente au cours de la journée.*

On estime que lorsque le nombre de clients dans la file d'attente est supérieur à 5 en moyenne sur la journée, il y a saturation. Pour $q = 0.1$, déterminer expérimentalement la valeur de p à partir de laquelle il y a saturation.

2.2 Seconde partie

Nous allons maintenant autoriser le bureau de poste à ouvrir plusieurs guichets.

Exercice 8. *Écrivez une fonction* `sim_day_2()` *qui prend trois arguments : les probabilités p et q et le nombre n de guichets ouverts. Elle devra aussi renvoyer le vecteur du nombre de clients présents dans la poste à chaque minute.*

Exercice 9. *Déterminez expérimentalement le nombre minimal de guichets requis pour ne pas atteindre le point de saturation ($5n$ clients en moyenne sur une journée), pour $q = 0.1$ et les valeurs de p suivantes :* $\{0.1, 0.25, 0.5, 0.75\}$.