Université Joseph Fourier **Frédérique Leblanc**
L2/STA230 – TP **leblanc@imag.fr**

TP3 : Simulation de données - Théorème Central Limite

Objectifs : *Utiliser les fonctions proposées par R pour simuler des données sous une loi usuelle (nouveautés ici : les lois de Bernoulli $\mathcal{B}(p)$ et Binômiale $\mathcal{B}(n,p)$) faire une description graphique et numérique des données simulées et la comparer avec la loi théorique utilisée pour effectuer la simulation. Illustrer le théorème central limite avec un modèle de Bernoulli et la loi des grands nombres avec un modèle normal.*

1 Loi Normale

Nous noterons ici X une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$ d'espérance (moyenne théorique) μ et variance (théorique) σ^2.

Exercice 1.1 :

Pour comprendre que plus on tire de réalisations d'une variable, plus la répartition observée (l'histogramme) et la moyenne empirique ($\bar{x}$) se rapprochent respectivement de la répartition théorique (densité de la loi simulée) et de la moyenne théorique (espérance de la loi simulée).

1. Créer deux objets sous R, appelés `mt` et `ett` qui recevront les valeurs 2 et 3.
2. Simuler un échantillon de taille $n = 10000$ de la variable X de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = 2$ et $\sigma = 3$ et le ranger dans `Ech` (rappel : le générateur aléatoire pour la loi normale est donné par la fonction `rnorm`).
3. Affecter dans l'objet `Ech1` (resp. `Ech2`, resp. `Ech3`, resp. `Ech4`, resp. `Ech5` resp. `Ech6`) les $n_1 = 100$ premières observations collectées dans `Ech` (resp. $n_2 = 200$, resp. $n_3 = 500$, resp. $n_4 = 1000$, resp. $n_5 = 5000$, resp. $n_6 = 10000$)
4. Partitionner la fenêtre graphique en six morceaux (3 lignes et 2 colonnes ou 2 lignes et 3 colonnes) et dans chaque morceau afficher successivement les histogrammes des 6 échantillons de tailles respectives $n_1 = 100$, $n_2 = 200$,..., $n_6 = 10000$ en renseignant la taille d'échantillon utilisée pour chaque graphe.
5. Ajouter sur chacun des histogrammes produits la densité de la loi normale de paramètre μ et σ. Conclusion ?
6. Revenez à une partition graphique comprenant un seul morceau, représentez-y la suite des moyennes empiriques de chacun des six échantillons et rajoutez-y la droite horizontale passant par le point $(0, 2)$. Conclusion ?

Exercice 1.2 :

Il s'agit ici d'étudier le comportement de $\bar{X}_n$ pour un échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$ d'espérance (moyenne théorique) $\mu = 2$, écart-type théorique $\sigma = 3$ et pour $n \geq 1$ puis de montrer que $\bar{X}_n$ suit aussi une loi normale dont on précisera les paramètres.

principe : Nous choisissons d'abord n, par exemple 2. Nous allons simuler $N = 1000$ réalisations de $\bar{X}_n$: pour cela nous avons besoin de $N = 1000$ réalisations de l'échantillon aléatoire de taille $n = 2$, (X_1, X_2). Pour chacune de ces 1000 réalisations du 2-échantillon (X_1, X_2) sera calculé la moyenne empirique $(\bar{X}_2)$. Nous disposerons alors d'un échantillon de $N = 1000$ réalisations de $\bar{X}_2$ pour lequel nous calculerons la moyenne empirique et la variance corrigée empirique et dont nous représenterons

la distribution avec un histogramme. Ensuite on y superposera la densité théorique d'une loi normale, convenablement choisie.

1. Créer les paramètres `N`, `n`, `mt`, `ett` et affectez leur les valeurs 1000, 2, 2 et 3.

2. Simuler 2000 réalisations de la loi $\mathcal{N}(\mu, \sigma^2)$ et rangez-les dans une matrice ayant 2 lignes et 1000 colonnes avec :
   `matrix(rnorm(2000,2,3),ncol=1000)->M`
   `dim(M)` # pour s'assurer que la matrice des données a les bonnes dimensions

3. Calculer les $N = 1000$ réalisations de la moyenne du 2-échantillon $\bar{X}_2$. On pourra utiliser la fonction `apply()` qui opère sur un tableau et permet d'appliquer le même calcul à toutes ses lignes ou colonnes. Ici on souhaite calculer les moyennes en colonnes (1000 colonnes) et les affecter à `Xbar` :
   `apply(M,MARGIN=2,mean) ->Xbar # calculs des moyennes en colonnes de M, si MARGIN=1 calculs en lignes.`
   Que produit la commande suivante : `apply(M,MARGIN=1,sd)` ?

4. Calculer la longueur (avec `length()`), la moyenne empirique (avec `mean()`), la variance corrigée s'^2 (avec `var()`), la variance empirique (avec `(length()-1)*var()/length()`) du vecteur des observations `Xbar`.

5. Représenter l'histogramme de ces données et y superposer la densité d'une loi $\mathcal{N}(\mu, \sigma^2/2)$ avec $\mu = 2$ et $\sigma = 3$ ainsi qu'une droite verticale passant par la moyenne empirique et une autre verticale passant par la moyenne théorique $\mu = 2$. Attention les arguments que demandent la fonction `dnorm()` sont successivement : la valeur de l'abscisse en laquelle on veut calculer la densité , le paramètre de centrage (ici c'est μ) et le paramètre de dispersion (ici c'est $\sqrt{\sigma^2/2}$). Conclusion ? Quelle est selon vous la loi que suit $\bar{X}_2$?

6. Qu'obtiendrait-on pour $n = 16$ (justifier) ?

2 Théorème Central Limite

Pour illustrer le théorème central limite qui dit que si on dispose d'un échantillon aléatoire $X_1, ..., X_n$ d'une même variable X de taille suffisante n, alors quelle que soit la loi de X choisie (nous choisirons la Bernoulli $\mathcal{B}(p)$ pour l'exercice numérique) la variable $\bar{X}_n = (\sum X_i)/n$ suit approximativement la loi $\mathcal{N}(E(X), V(X)/n)$ (où $E(X)$ et $V(X)$ désignent respectivement l'espérance et la variance de la variable X). Rappel : sous un modèle de Bernoulli $E(X) = p$ et $V(X) = p(1-p)$.

Exercice 2.1 :
On considère ici que la variable générant les échantillons est X de loi $\mathcal{B}(p)$.

1. Créer le paramètre `p` dans le quel vous affecterez la valeur 0.05. Affecter également p dans `mt` et $\sqrt{p * (1-p)}$ dans `ett`

2. Créer une matrice `B` ayant $n = 2$ lignes et $N = 1000$ colonnes et dont les éléments sont des réalisations d'une variable de Bernoulli de paramètre $p = 0.05$. La Bernoulli étant une binômiale particulière on utilisera le générateur aléatoire d'une binômiale donné par la fonction `rbinom()`. En consultant l'aide en ligne de la fonction vous apprendrez les arguments attendus et l'ordre dans lequel les donner (pour information : les fonctions `dbinom()`, `pbinom()` sont la densité et la fonction de répartition de la loi binômiale).

3. En appliquant les mêmes lignes de commande que pour réaliser les questions 4 et 5 du précédent exercice, (mais avec `B` au lieu de `M`) représentez l'histogrammes des observations obtenues de $\bar{X}_2$ et calculez les principaux résumés numériques. Peut-on modéliser la loi de $\bar{X}_2$ par une loi $\mathcal{N}(E(X), V(X)/2)$ avec $E(X) = p$ et $V(X) = p(1-p)$? ... Pourquoi non ?

4. Refaites les questions précédentes pour un n-échantillon pour lequel on prend $n = 200$ (On effectuera aussi $N = 1000$ réalisations de l'échantillon aléatoire $(X_1, ..., X_{200})$ que l'on rangera dans un tableau à 200 lignes et 1000 colonnes). Conclusion ?