MST - Logiciel R

TP n° 3

Distribution d'échantillonnage, estimation

Exercice n°1 – Quelques fonctions R utiles pour générer des données

1. La commande `i:j` permet de générer tous les entiers entre `i` et `j`.
   - Créer le vecteur comprenant les entiers entre 1 et 10.
   - Construire le vecteur de dimension 10 où l'élément i est la somme des i premiers entiers (utiliser la fonction `cumsum`.)
2. La fonction `rep(x,n)` permet de générer l'objet `x` (un nombre, un vecteur) `n` fois.
   - Créer un vecteur de taille 10 composé uniquement de "1"
   - Créer le vecteur `(1,2,3,1,2,3,1,2,3,1,2,3)`
   - Créer la matrice $\begin{pmatrix} 1 & 3 & 4 & 4 & 5 \\ 2 & 3 & 4 & 5 & 5 \\ 2 & 3 & 4 & 5 & 5 \end{pmatrix}$
   - Créer un jeu de 32 cartes (du 7 à l'As) comportant les quatre couleurs cœur, carreau, pique et trèfle (indication : utiliser la fonction `paste` permettant de concaténer des chaînes de caractère).
3. La fonction `seq (from=a, to=z, by=p)` permet de générer une séquence allant de `a` jusqu'à `z` avec un pas `p`.
   - Créer le vecteur des entiers de 1 à 10.
   - Construire le vecteur `(5.0,5.5,6.0,6.5,7.0,7.5,8.0,8.5,9.0,9.5,10.0)`
   - Construire le vecteur `(1.0,0.9,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1,0.0)`
   - Tracer la fonction `sin(x)` avec `x` variant entre $-\pi$ et $+\pi$ avec un pas de `0.01`.
4. La fonction `sample(e,n,replace=TRUE or FALSE)` permet d'extraire un échantillon de taille `n` d'un ensemble fini `e`, avec ou sans remise.
   - Simuler un échantillon de 60 lancers d'un dé équilibré. Calculer les fréquences de chaque numéro à l'aide de la fonction `table`.
   - Prendre au hasard 8 cartes sans remise dans le jeu de 32 cartes construit plus haut. Quelle est la probabilité d'obtenir exactement un As dans la main obtenue ?

Exercice n°2 – Échantillons aléatoires, calcul de probabilité et de quantile

Le logiciel R permet de générer aléatoirement des échantillons issus de nombreuses lois de probabilité discrètes ou continues (voir Table 1). En pratique, on utilise ces lois avec les préfixes suivants :

`d` : pour obtenir la valeur de la densité de probabilité en un point (ex. `dnorm`).
`q` : pour calculer un quantile à partir d'une probabilité cumulée (ex. `qpois`).
`p` : pour la fonction de répartition en un point (ex. `phyper`).
`r` : pour générer un échantillon aléatoire (ex. `runif`).

Ces lois, leurs différentes utilisations ainsi que leur représentation graphique sont accessibles dans le menu *Distribution* de `Rcmdr`.

1. Calculer la valeur de la densité normale centrée réduite au point 0.
2. Produire un échantillon de taille $n = 30$ issu d'une loi binomiale de paramètres $n = 10$ et $p = 0.5$.
3. Quel est le quantile d'ordre 0.975 d'une loi de Student à 20 degrés de liberté ? D'ordre 0.025 ?
4. Quelle est la probabilité d'avoir exactement un as dans 8 cartes extraites au hasard d'un jeu de 32 cartes ?
5. On suppose que la taille d'une femme est distribuée selon une loi normale de paramètre $\mu = 163$ et $\sigma = 6$. Quelle est la probabilité de trouver une femme de plus de deux mètres dans cette population ? Une femme dont la taille est comprise entre 150 et 155 cm ? Quelle est la taille la plus petite des 10% des femmes les plus grandes ?

Exercice n°3 – Convergence d'une fréquence de pile ou face

1. Générer un échantillon de taille $n = 100$ simulant les résultats de 100 jets d'une pièce de monnaie équilibrée (utilisation de la commande `rbinom` en langage de commande ou dans le cadre du menu *Distribution* de Rcmdr).
2. À l'aide de la fonction `cumsum`, créer le vecteur où l'élément i est l'effectif cumulé du nombre de succès après i essais.
3. Créer le vecteur $(1, 2, ..., n)$ à l'aide de la fonction `seq`. En déduire le vecteur des fréquences cumulées de succès après i essais.
4. Tracer le graphique avec en abscisse le nombre d'essais et en ordonnée la fréquence cumulée de succès. Ajouter la ligne horizontale au niveau de la probabilité théorique de succès (utiliser la commande `abline`.)

Exercice n°4 – Simulation d'échantillons - Échantillonnage

1. À l'aide du menu *Distribution* de Rcmdr, simuler 1000 échantillons (en lignes) de taille 30 (en colonnes) issus d'une loi uniforme continue sur l'intervalle $[0; 1]$. Paramètres : nommer le jeu `data1` et ajouter la moyenne et l'écart type de l'échantillon au jeu de données. Visualiser les données obtenues.
2. Ajouter la variance d'échantillon. Deux possibilités :
   – Par le langage de commande : `data1$var = data1$sd^2`
   – Par le menu : *Données* → *Gérer les variables* → *Calculer une nouvelle variable*. Nommer la variable `var` et entrer l'expression `sd^2`.
3. Produire les statistiques descriptives (moyenne, écart type) pour toutes les colonnes du tableau de données obtenu.
4. À quoi est égale la moyenne des moyennes d'échantillon ? De quelle valeur théorique cette moyenne empirique est-elle une estimation ?
5. À quoi est égal l'écart type des moyennes d'échantillon ? De quelle valeur théorique cet écart type empirique est-il une estimation ?

6. Tracer l'histogramme des valeurs de la première colonne du tableau obtenu. Que représente ce graphique ?
7. Représenter l'histogramme des valeurs de la moyenne d'échantillon (colonne `mean`). Demander un nombre de classes élevé (par ex. 35) et choisir la densité en ordonnée. Que constate-t-on ? Vers quelle loi converge la distribution de la moyenne lorsque le nombre d'observations tend vers l'infini ? Superposer au graphique une densité normale appropriée (fonction `curve`).
8. Tracer l'histogramme des valeurs de la variance d'échantillon. Commenter.
9. Reprendre la construction de l'histogramme de la moyenne d'échantillon (pour 1000 tirages d'échantillons issus d'une loi uniforme) lorsque les échantillons sont de taille $n = 3$. Que constate-t-on ?

Exercice n°5 – Calcul de vraisemblance

1. L'échantillon ci-dessous donne le résultat de 10 lancers d'une pièce de monnaie dont la probabilité de *pile* (associé à la valeur 1) est supposée égale à P :
   ```
   > ech = c(1,1,1,0,1,1,1,0,0,1)
   ```
   Calculer la distribution de fréquences pour cet échantillon.
2. Écrire la vraisemblance de cet échantillon comme une fonction du paramètre P.
3. Calculer la vraisemblance pour des valeurs de P variant de 0 à 1 avec un pas de $\frac{1}{100}$. Tracer la valeur de la vraisemblance en fonction des valeurs de P.
4. Pour quelle valeur de P la vraisemblance est-elle maximale ?

Exercice n°6 – Intervalle de confiance

1. Importer le fichier `superficie.txt` qui fournit la superficie en kilomètres carrés des 96 départements de la France métropolitaine. Nommer le jeu de données `pop`.
2. Calculer la superficie moyenne d'un département ainsi que la superficie totale de la France.
3. À l'aide de la fonction `sample`, tirer aléatoirement dix numéros de départements entre 1 et 96.
4. Construire le jeu de données correspondant à l'échantillon des 10 superficies obtenues. Nommer `ech10` ce jeu de données.
5. Calculer la moyenne et l'écart type corrigé de l'échantillon.
6. Construire l'intervalle de confiance au seuil de 95% pour la superficie moyenne d'un département français.
   Remarques :
   - Attention, la population totale étudiée est de taille finie N !
   - Quelle hypothèse concernant la distribution de la superficie d'un département est-il nécessaire de poser ? Qu'en pensez-vous ?
7. Votre intervalle de confiance contient-il la vraie superficie moyenne d'un département ? À quel résultat peut-on s'attendre pour l'ensemble des intervalles de confiance construits par les étudiants de ce TP ?
8. Quel est l'intervalle de confiance pour la superficie totale de la France ?

Loi	Fonction R	Arguments
bêta	`beta`	forme 1, forme 2
binomiale	`binom`	size, prob
chi deux	`chisq`	df (degrés de liberté)
uniforme	`unif`	min, max
exponentielle	`exp`	rate
Fisher	`f`	df1, df2
gamma	`gamma`	forme, échelle
géométrique	`geom`	prob
hypergéométrique	`hyper`	m, n, k (taille échantillon)
binomiale négative	`nbinom`	size, prob
normale	`norm`	mean, sd
Poisson	`pois`	lambda
Student	`t`	df
Weibull	`weibull`	forme, échelle

TABLE 1 – Principales lois de probabilités et fonction R associées.