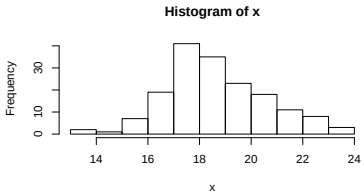
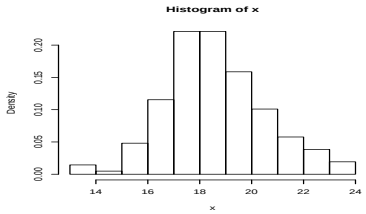
Cours 01
Histogrammes

1 Individus et caractères

On appellera *échantillon* de *taille* n un ensemble d'*individus* i que, par commodité, on nomme $1, 2, 3, \ldots, n$. Ces individus peuvent être des personnes, ou des objets, ou des pays... Chaque individu présente des *caractères*, notés x, y, etc. Un caractère est généralement une mesure x_i telle la taille (en cm) de l'individu i, ou la mesure y_i de son poids (en kg). Dans ce cas on dit que le caractère est *numérique* (ou *continu*). Un caractère peut aussi être *catégorique* (ou *qualitatif* ou *discret*) : sexe (M ou F), couleur des yeux (noire, marron, bleus). Nous considèrerons surtout le cas de caractères numériques car ils se prêtent à des calculs tels la moyenne ou l'écart-type. La commande `library(MASS)` de `R` donne accès à un grand nombre d'échantillons et la commande `data()` en donne les noms.

2 Histogrammes





L'histogramme d'un caractère x est un résumé graphique des valeurs que prend ce caractère sur un échantillon. Il comporte K *classes* C_k qui sont des intervalles consécutifs $C_k =]c_{k-1}, c_k]$ tels que le caractère x_i de chaque individu i soit dans l'un de ces intervalles. On compte alors le nombre e_k d'individus i dont le caractère x_i est dans la classe C_k ; ce nombre e_k s'appelle l'*effectif*[1] de la classe C_k. L'histogramme du caractère x représente alors un barreau de hauteur e_k au-dessus de l'intervalle $]c_{k-1}, c_k]$. En d'autres termes, c'est le graphe de la fonction $x \mapsto e(x) = \sum_{k=1}^{K} e_k I_{C_k}(x)$, où I_{C_k} désigne la fonction indicatrice de l'intervalle C_k, c'est-à-dire que $I_{C_k}(x)$ est égal à 1 ou 0 selon que $x \in C_k$ ou $x \notin C_k$. Les effectifs des classes de la figure de gauche ci-dessus vont de $e_2 = 1$ pour la classe $c_2 =]14, 15]$ à $e_5 = 53$ pour la classe $c_5 =]17, 18]$. L'histogramme présente $K = 11$ classes.

3 Tracé d'histogrammes au moyen de R

Nous avons vu en cours comment lire un fichier de données et placer dans un *data-frame* de `R` les valeurs de différents caractères pour les divers individus de l'échantillon constitué par le fichier. La commande `hist(x)` permet de représenter, dans une nouvelle fenêtre, la représentation graphique d'un histogramme d'un caractère x pour l'échantillon considéré. Il est important d'observer que, ce faisant, `R` effectue de nombreux choix à notre place. En effet, il y a de nombreux histogrammes possible. Le premier choix à faire est le nombre K de classes. Un choix souvent raisonnable est la partie entière de $\sqrt{K}$, mais il y a aussi de bonnes raisons pour choisir $K = 1 + \frac{10}{3}\log(n)\ldots$. Ensuite il faut choisir les bornes c_0, …, c_K. Un choix naturel peut-être $c_0 =$`min(x)`, $c_K =$`max(x)`, puis de choisir des valeurs de c_k également espacées. Toutefois, ceci a peu de chance d'aboutir à des valeurs simples pour le c_k, ce qui fait que `R` procède à des choix un peu différents, pour aboutir à des bornes si possible entières. Nous verrons en exercice comment imposer d'autres bornes que celles choisies par `R`. L'écart entre la plus petite valeur `min(x)` de l'échantillon et sa plus grande valeur `max(x)` s'appelle l'*étendue* de l'échantillon. La commande `hist(x,freq=FALSE)` permet de représenter un histogramme semblable, mais cette fois ce ne sont par les effectifs qui donnent la hauteur des barres mais la fraction du total n, $\frac{e_k}{n}$. Dans ce cas `R` remplace alors son `frequency` par `density`, comme dans la figure de droite ci-dessus. Cela revient à remplacer la fonction $x \mapsto e(x)$ par la fonction $x \mapsto h(x) = \sum_{k=1}^{K} h_k I_{C_k}(x)$ telle que $(c_k - c_{k-1}) * h_k = \int_{c_{k-1}}^{c_k} h(x)dx = e_k/n$, et donc $\int_{\min(x)}^{\max(x)} h(x)dx = 1$.

1. appelé de façon impropre `frequency` par `R`.

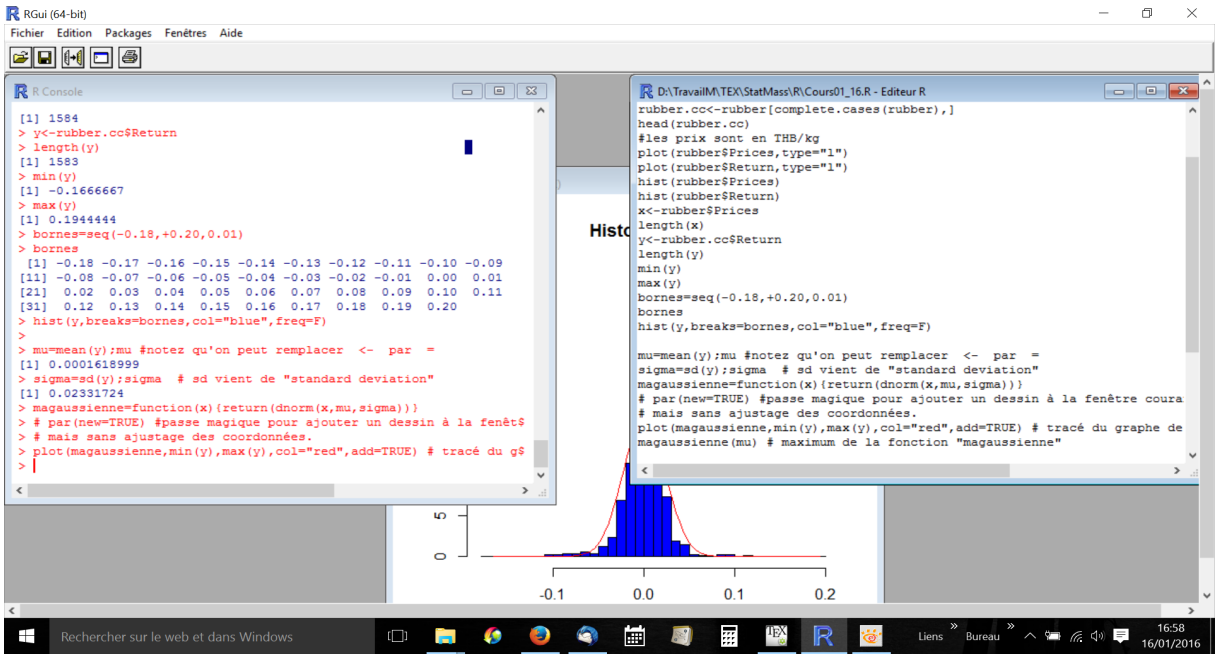


FIGURE 1 – Au lancement de R, une première fenêtre s'ouvre : il s'agit de la console. C'est là que s'afficheront les résultats numériques, mais aussi les messages d'erreur : il convient donc qu'elle soit en permanence visible. Évitez de l'utiliser pour passer des commandes : pour celles-ci, ouvrez un éditeur et fixez les tailles de deux fenêtres comme ci-dessus. Pour exécuter une ligne de l'éditeur placez votre curseur dans cette lignes et cliquez sur le bouton comportant une grosse flèche de gauche à droite. Pensez à sauver régulièrement votre éditeur, par *Edition-Sauver*, (ou *Ctrl-S*). Commencez par nommer le fichier de sauvegarde, par *Edition-Sauver sous...*. Placez des commentaires (après le caractère `#`)

Quelques instructions utilisées

`read.csv2()` : instruction de lecture dans un fichier où les nombres comportent des virgules décimales et où les séparateurs sont de point-virgules.

`file.choose` : instruction ouvrant une fenêtre-système permettant la navigation pour choisir un fichier

`<-` : intruction d'affectation ; peut se remplacer par `=`

`complete.case()` : cette instruction permet de déterminer quelles sont les lignes d'un data-frame qui sont complète, c'est-à-dire qui ne comporte pas de `NA`, la désignation d'une donnée manquante.

`rubber$Prices` : la variable colonne de nom `Prices` du *data-frame* `rubber`

`seq(-0.18,+0.20,0.01)` : la suite des valeurs -0.18, ..., $+0.20$.

`hist()` : instruction de tracer d'un histogramme. Par défaut `freq=T` et produit un histogramme en effectifs. L'option `freq=F` permet de remplacer le graphe en $x \mapsto e(x)$ par celui de $x \mapsto h(x)$. Cette fonction peut se comparer à une fonction de densité théorique, ici une densité gaussienne.

Le fichier rubber.csv

Nous utilisons les données des prix du caoutchouc naturel (latex) observés sur le marché de Hat Yai situé au cœur la région productrice de Songkhla au Sud de la Thailande. Ces prix P_t sont exprimés en Bath thaïlandais THB par kilogramme. Nous pourrons les comparer avec les prix observés sur le marché de matière première de Tokyo, le TOCOM. Ce sont les rendements $R_t = (P_t - P_{t-\delta t})/P_{t-\delta t}$ qui se prêtent bien à un modèle gaussien.