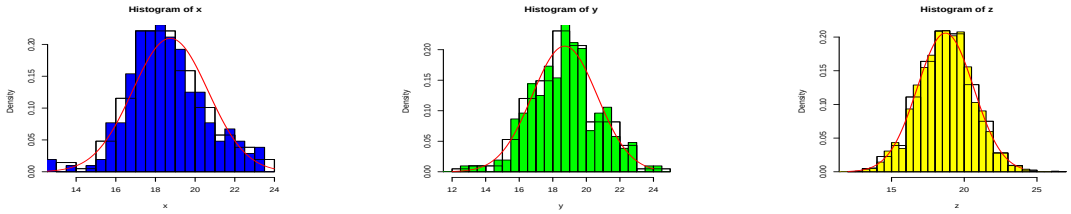
Cours 02
Fonction de répartition et densité sur l'exemple de la loi gaussienne

1 Densité

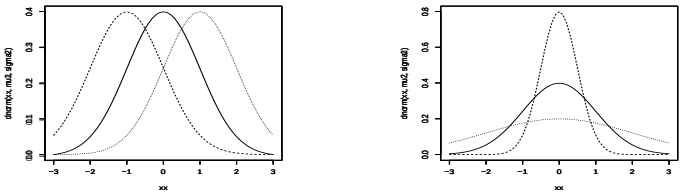


La notion d'histogramme d'un échantillon donne une représentation graphique des valeurs d'un caractère qui dépend des bornes choisies pour les classes. Des choix différents donnent des histogramme bien différents, comme cela peut se vérifier dans les dessins ci-dessus où l'on a simplement choisi de subdiviser en deux les classes entre les deux histogrammes superposés sur un même dessin. Un point commun aux deux histogrammes est la forme "en cloche" (avec d'inévitables "créneaux"). Dans d'autres situations, on pourrait obtenir d'autres formes, mais nous allons aborder ici un modèle continu (ou "lisse") pour cette forme en cloche, dit *modèle gaussien*. La pertinence d'un type de modèle est une question importante en statistique et débouche naturellement sur celle de test que nous étudierons plus tard. Notons qu'en principe l'inconvénient des créneaux diminue avec la taille de l'échantillon : c'est ce que nous avons illustré entre le dessin du milieu et celui de droite : ils correspondent à deux échantillons produits de manière similaire (par "simulation"), sauf que le dessin de droite correspond à un échantillon dix fois plus gros que celui du dessin du milieu (et celui de gauche qui est le même que celui vu au cours précèdent). A noter que nous avons opté pour une représentation où la hauteur des barres est la fraction $\frac{e_k}{n}$ de l'effectif total n. Ainsi la somme des aires des barres est égale à 1. On dira que l'histogramme est *normalisé*. On cherche alors la ressemblance de l'histogramme avec une fonction de densité $x \mapsto f(x)$, c'est-à-dire une fonction positive et d'intégrale $\int_{-\infty}^{+\infty} f(x)dx = 1$, donc délimitant, elle aussi, une région d'aire égale à 1.

Si `x` désigne l'échantillon considéré, la figure de gauche a été obtenue au moyen du code suivant :

```
hist(x,freq=F)
xmin<- floor(min(x))-1;xmax<- ceiling(max(x))
pas=0.5
bornes=seq(xmin,xmax,pas);bornes
hist(x,breaks=bornes,add=TRUE,col="blue",freq=F) # notez les rôles de "add" et "col"
mu=mean(x);mu # notez qu'on peut remplacer <- par =
sigma=sd(x);sigma # sd vient de "standard deviation"
magaussienne=function(x){return(dnorm(x,mu,sigma))} # voir ci-dessous pour dnorm
plot(magaussienne,xmin,xmax,col="red",add=TRUE) # tracé du graphe de la fonction
```

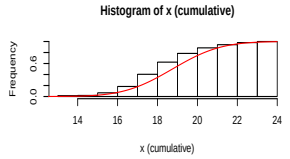
2 Densité gaussienne

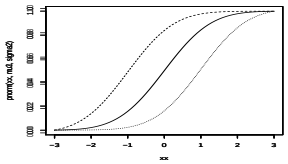


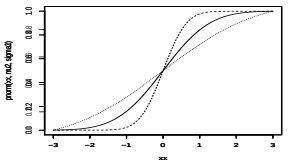
On appelle *densité gaussienne* (ou simplement *gaussienne*) de paramètre μ et $\sigma > 0$ la fonction $t \mapsto f_{\mu,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$. Son graphe est symétrique autour de la valeur μ puisque $f_{\mu,\sigma}(\mu+t) = f_{\mu,\sigma}(\mu-t)$. Si $\mu = 0$ on dit que la gaussienne est *centrée*. Si $\sigma = 1$ on dit que la gaussienne est *réduite*. Notons que

plus σ est petit, plus la cloche est étroite ; mais comme l'aire qu'elle limite est toujours égale à 1, plus σ est petit, plus la valeur maximale $f_{\mu,\sigma}(\mu)$ est grande.

Notons $F_{\mu,\sigma}(x) = \int_{-\infty}^{x} f_{\mu,\sigma}(t)dt$ la primitive de $f_{\mu,\sigma}$ telle que $\lim_{x\to-\infty} F_{\mu,\sigma}(x) = 0$; alors $\lim_{x\to+\infty} F_{\mu,\sigma}(x) = 1$ car $f_{\mu,\sigma}$ est une densité. Cette fonction s'appelle la *fonction de répartition* de la loi gaussienne. Elle modélise l'*histogramme cumulatif* (croissant) d'une distribution en cloche.







La loi gaussienne est également appelée *loi normale* ; c'est pourquoi les fonctions de `R` liées à la gaussienne ont un nom comportant la racine `norm`. Ainsi la densité gaussienne est appelée `dnorm` ; pour `mu`$= \mu$ et `sigma`$= \sigma$, $f_{\mu,\sigma}(x)$ =`dnorm(x,mu,sigma)` où le "`d`" rappelle qu'il s'agit de la densité. De même $F_{\mu,\sigma}(x)$ =`pnorm(x,mu,sigma)` où le "`p`" rappelle qu'il s'agit de la probabilité d'être inférieur à x. L'histogramme cumulatif ci-dessus a été obtenu au moyen du code suivant

```
hist(x,freq=F)->h #histogramme des valeurs de x
h$counts <- cumsum(h$counts)/length(x) # remplace freq.s par cumulative freq.s
h$xname <- "x (cumulative)"
plot( h ) # l'histogramme est à présent cumulatif
maGaussienne=function(x){return(pnorm(x,mu,sigma))}
plot(maGaussienne,xmin,xmax,col="red",add=TRUE)
```

3 Choix de μ et de σ

On peut montrer que $\mu = \int_{-\infty}^{+\infty} t f_{\mu,\sigma}(t)dt$ et que $\sigma^2 = \int_{-\infty}^{+\infty} (t-\mu)^2 f_{\mu,\sigma}(t)dt$. Si l'on interprète l'intégrale comme une version lisse d'une somme et $f_{\mu,\sigma}(t)dt$ comme l'aire d'une barre d'histogramme d'une classe étroite (de largeur dt) autour de la valeur t on voit qu'il est naturel de choisir μ comme étant la moyenne $\hat{\mu}$ des valeurs x_i du caractère x. De même il convient de choisir σ^2 comme la moyenne des $(x_i - \mu)^2$ et donc σ comme la racine $\hat{\sigma}$ de la moyenne des $(x_i - \mu)^2$. Les nombres $\hat{\mu}$ et $\hat{\sigma}$ s'appellent la *moyenne* et l'*écart-type* de l'échantillon x. Ils sont calculés par les commande `mean(x)` (moyenne, en anglais) et `sd(x)` (standard deviation, écart-type en anglais). Nous reviendrons sur la notion générale de moyenne et d'écart-type.

4 Retour à la modélisation des rendements observés

Sur le diagramme ci-dessous nous avons reproduit l'histogramme en densités des valeurs observées du rendement $R_t = (S_t - S_{t-\delta t})/S_{t-\delta t}$ sur les prix S_t observés sur le marché du latex de Hat Yai. Il a été obtenu au moyen de l'instruction

`hist(y,breaks=bornes,col="blue",freq=F)` avec

`bornes=seq(-0.18,+0.20,0.01)`.

Nous avons superposé le graphe de la gaussienne ajustée aux données en adoptant la moyenne `mean(y)` et l'écart-type `sd(y)` de l'échantillon `y` de rendements en guise de μ et σ pour la gaussienne. Ceci a été obtenu par `plot(magaussienne,min(y),max(y),col="red",add=TRUE)` avec

`magaussienne=function(x)return(dnorm(x,mu,sigma))` . Cette image indique (au moins) deux défauts pour un modèle "purement" gaussien : la classe d'effectif maximal s'écarte largement du modèle gaussien adopté. L'examen des données montre que ceci est du au nombre excessif de jours où les prix restent inchangés par rapport à ceux de la veille. Un autre défaut, moins visible à première vue, est la présence excessive de rendements élevés en valeur absolue. ces deux constats peuvent suggérer de rechercher un modèle plus élaboré, en recherchant les causes possibles pour ces deux observations statiques.