Centre universitaire Abedelhafid Boussouf – Mila
Institut des sciences et technologie
Département de mathématiques et informatique
1er année Master : Intelligence artificielle et ses applications – Année: 2023-2024
Matière : Modélisation et Simulation
Responsable de la matière : Dr. Sadek Benhammada

Table des matières

Ajustement de la distribution théorique (Theoretical distribution fitting)

L'ajustement de la distribution est un processus classique d'estimation statistique qui consiste à sélectionner une distribution de probabilité théorique pour modéliser des données observées, à estimer les paramètres de cette distribution, et à évaluer la qualité de l'ajustement.

Ce processus est d'une importance cruciale dans la modélisation des temps d'arrivée des clients et des temps de service dans les files d'attente. L'objectif est de représenter de manière précise la nature stochastique de ces processus, permettant ainsi aux modèles de simulation de refléter fidèlement les scénarios du monde réel.

Dans ce tutoriel, nous présentons les différentes étapes de l'ajustement de distributions théoriques à l'aide du package **fitdistrplus**. Ce package implémente plusieurs fonctions dédiées pour faciliter ce processus.

Pour installer le package fitdistrplus dans R, il suffit de taper dans le console :

```
install.packages("fitdistrplus")
```

1 Vérification de l'indépendance des données

La première étape de l'ajustement de la distribution théorique consiste à vérifier si les données collectées sont indépendantes. En effet, l'ajustement de distribution suppose que les données sont indépendantes les unes des autres. Si les données ne sont pas indépendantes, l'ajustement de distribution peut produire des résultats erronés.

Différentes méthodes peuvent être utilisées pour vérifier l'indépendance de données :

1.1 Diagramme de dispersion

Une méthode simple pour évaluer l'indépendance des données consiste à tracer un diagramme de dispersion. Pour les données $X_1, X_2 \dots X_n$ répertoriés dans l'ordre temporel de la collecte, les paires (X_i, X_{i+1}) pour $i = 1 \sim n - 1$ sont tracées sur un système de coordonnées (X_i comme valeur de x et X_{i+1}comme valeur de y).

- Si les points tracés sont dispersés au hasard, on peut conclure que les données sont indépendantes.
- Si les points semblent alignés, cela peut suggérer une dépendance entre les données.

```
# Code R : Diagramme de dispersion
data <-...
# Tracer un diagramme de dispersion de X_i et X_i+1
plot(data [-1], data [-length(data)])
```

1.2 Fonction d'autocorrélation

Une méthode pour vérifier l'indépendance de données consiste à utiliser la fonction **acf()** (autocorrelation function) pour visualiser l'autocorrélation dans les données. Si l'autocorrélation chute rapidement vers zéro, cela suggère une faible dépendance.

```
# Code R : Fonction de d'auto-correlation
data <-...
# Fonction d'auto-correlation pour verifier l'independance
acf(data)
```

1.3 Coefficients de corrélation de Pearson et Spearman :

Les coefficients de de Pearson et de Spearman prennent des valeurs de -1 à 1.

- Plus les valeurs sont proches de 0, plus il y a une indication d'indépendance.
- Une valeur proche de 1 ou de -1 indique une dépendance entre les données.

```
# Code R: Coefficients de corrélation
data <-...
# Coefficients de corrélation de Pearson et Spearman
cor(data [-1], data [-length(data)], method="pearson")
cor(data [-1], data [-length(data)], method="spearman")
```

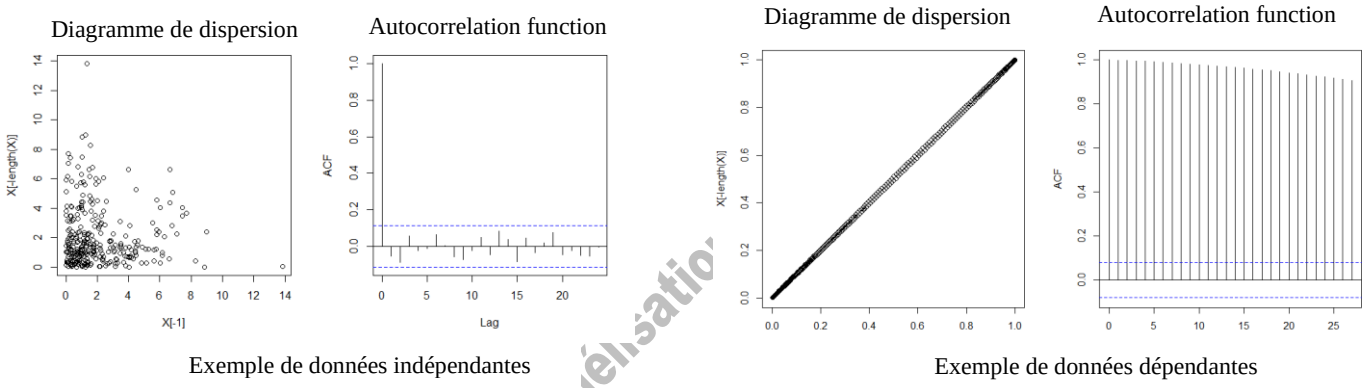


Figure 1. Exemple de données indépendantes (à gauche) et dépendantes (à droite)

2 Sélection de la fonction de distribution

La deuxième étape consiste à sélectionner une distribution candidate appropriée. Ce choix peut être guidé par :

- La connaissance des processus stochastiques régissant la variable modélisée. Par exemple, la distribution exponentielle et la distribution d'Erlang sont généralement sélectionnées pour les temps inter-arrivées, tandis que les distributions de temps de service largement utilisées sont la distribution bêta et la distribution log-normale. Le tableau 2 récapitule les applications possibles des lois de probabilités théoriques.
- L'observation des propriétés statistiques (moyenne, écart type, skewness, kurtosis, etc.), les graphiques de densité de probabilités et de la fonction de répartition empirique.

2.1 Fonction `plotdist`

Les tracés de la fonction de distribution empirique et de densité de probabilités, qui peuvent être obtenus avec la fonction **`plotdist`** du package **`fitdistrplus`.**

Cette fonction fournit deux tracés (voir Figure 4.5) : le tracé de gauche est l'histogramme sur une échelle de densité (et/ou le tracé de densité des deux, en fonction des valeurs des arguments **`histo`** et **`demp`**) et le tracé de droite est la fonction de distribution cumulée empirique (CDF).

```
# Code R
library(fitdistrplus)
data <- ...
plotdist(data, histo = TRUE, demp = TRUE)
```

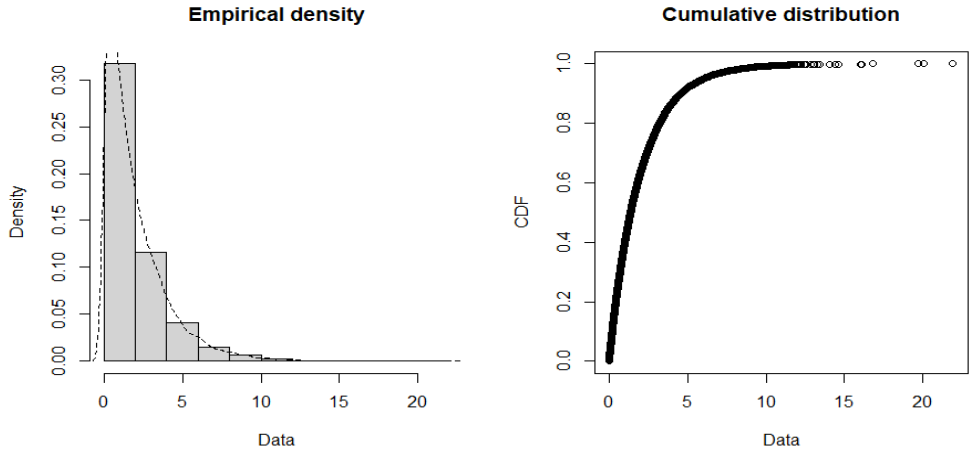


Figure 2. Densité de probabilités et tracés CDF empirique d'un échantillon de données

2.2 Fonction `descdist`

En plus des tracés empiriques, les statistiques descriptives peuvent aider à choisir des candidats pour décrire une distribution parmi un ensemble de distributions paramétriques. L'**asymétrie** (**skewness**) et l'**aplatissement** (**kurtosis**) sont particulièrement utiles à cet effet.

La fonction **`descdist`** du package **`fitdistrplus`** fournit des statistiques descriptives classiques (minimum, maximum, médiane, moyenne, écart type), d'asymétrie (**skewness**) et d'aplatissement (**kurtosis**) :

```
Code R : Fonctions plot et descdist
library(fitdistrplus )
data <- ...
plotdist(data, histo = TRUE, demp = TRUE)
descdist(data, boot = 1000)
```

```
summary statistics
------
min:  0.000773102   max:  17.85025
median:  1.382162
mean:  1.994817
estimated sd:  2.003813
estimated skewness:  2.047466
estimated kurtosis:  9.283356
```

Un tracé d'asymétrie-aplatissement (skewness-kurtosis" est fourni par la fonction **descdist** de distribution empirique (Cullen and Frey graph - voir la figure 4.6).

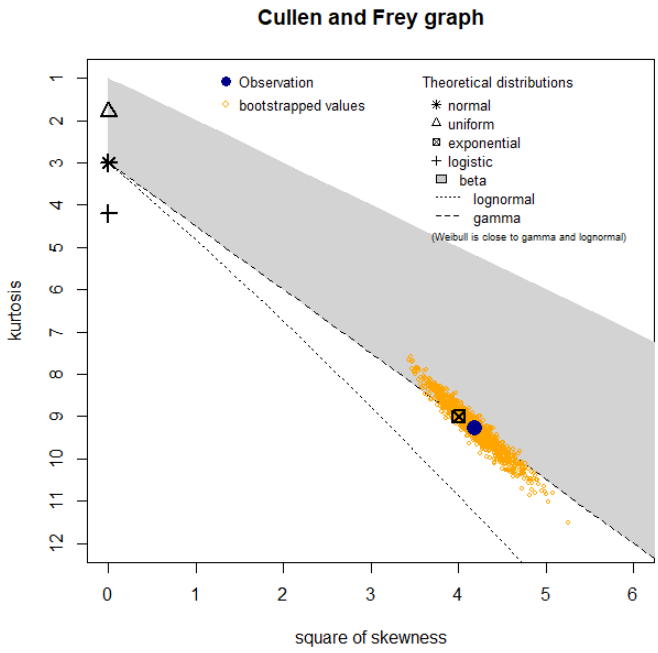


Figure 3. Densité de probabilités et tracés CDF empirique d'un échantillon de données

- Sur le graphique, les valeurs des distributions courantes sont affichées afin de faciliter le choix des distributions à adapter aux données.
- Pour certaines distributions (normale, uniforme, logistique, exponentielle), il n'y a qu'une seule valeur possible pour l'asymétrie et l'aplatissement. Ainsi, la distribution est représentée par un seul point sur le tracé.
- Pour les autres distributions, des zones de valeurs possibles sont représentées, constituées de lignes (comme pour les distributions gamma et lognormale) ou de zones plus grandes (comme pour la distribution bêta).
- Afin de prendre en compte l'incertitude des valeurs estimées d'aplatissement et d'asymétrie à partir des données, une procédure bootstrap peut être effectuée en utilisant l'argument **`boot`**. Les valeurs d'asymétrie et d'aplatissement sont calculées sur des échantillons bootstrap (construits par échantillonnage aléatoire avec remplacement à partir de l'ensemble de données d'origine) et rapportées sur le graphique d'asymétrie-aplatissement (points jaunes).
- En raison de l'incertitude des valeurs estimées d'aplatissement et d'asymétrie le tracé asymétrie-aplatissement doit être considéré comme étant uniquement **indicatif**.

Distribution	Skewness	Pearson's kurtosis
Normale	0	3
Uniforme	0	1.8
Exponentielle	2	9

Tableau 1. Skewness et Pearson's kurtosis des distributions Normale, Uniforme, et Exponentielle

Le choix de la distribution théorique est appuyé par les éléments suivants :

- La comparaison des graphiques de densité de probabilités et de la fonction de répartition empirique à l'aide de la fonction **`plotdist`** (Figures 2), avec les graphiques de la fonction de densités et de la fonction de répartition des lois théoriques.
- L'examen des statistiques descriptives des données fournies et représentées graphiquement à l'aide la fonction **`descdist,`** en les comparant à celles des lois théoriques (Figures 3).
- La prise en compte des applications possibles des distributions théoriques (Tableau 2).

Exemple

- En comparant les formes des densités de probabilités et des fonctions de répartition (Figures 4.5), ainsi que les statistiques descriptives (Figures 4.6), les distributions qui pourraient être considérées sont : Exponentielle, Beta, Gamma (Erlang) et Weibull.
- Si les données représentées sont des temps inter-arrivées, le choix de la distribution théorique se fera entre les lois Exponentielle, Erlang et Weibull, appliquées pour modéliser les temps inter-arrivées (Tableau 4.3).
- Si les données représentées sont des temps de service, la distribution théorique adéquate est la loi Beta.

Distributions	**Applications**
$\boldsymbol{Exponentielle\ (\theta)}$	Modélisation des **temps inter-arrivées** des « clients » qui se produisent à un taux constant ; Modélisation des temps inter-défaillances d'un équipement (interfailure times).
$\boldsymbol{Erlang\ (k, \theta)}$	Modélisation des **temps inter-arrivées** des « clients » ;
Weibull($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$)	Modélisation des **temps inter-arrivées des** « clients » qui se produisent à un taux qui augmente ou diminue dans le temps ; Modélisation des temps inter-défaillances d'un équipement (interfailure times). – Lorsque $\beta > 1$, le taux augmente avec le temps (phase d'usure), – $\beta = 1$ correspond à un taux constant (distribution exponentielle), – Lorsque $0 < \beta < 1$, le taux de diminue avec le temps.
$\boldsymbol{Uniforme(a, b)}$	Utilisée comme « premier » modèle pour une quantité qui semble varier de manière aléatoire entre *a* et *b* mais sur laquelle on a peu de données. Modélisation des **temps de services** lorsque seule la plage [a, b] des durées de service est fournie.
$\boldsymbol{Triangulaire\ (a, b, c)}$	Modélisation des **temps de services** si le mode c est également donné en plus de la plage [a, b].
$\boldsymbol{Bêta\ (\alpha, \beta)}$	Modélisation des **temps de services** avec une plage finie. La distribution bêta standard $Y \sim Beta(\alpha, \beta)$ a une plage unitaire [0, 1]. Alors la variable aléatoire bêta X avec une plage générale $[a, b]$ peut être obtenue à partir de Y comme suit : $X = a + Y(b - a)$
$\boldsymbol{Normale\ (\mu, \sigma)}$	Si la distribution des **temps de services** est symétrique vers la droite (asymétrie nul), ils sont générés à partir de la distribution log-normale.
$\boldsymbol{Lognormal\ (\mu, \sigma)}$	Si la distribution des **temps de services** est asymétrique vers la droite (asymétrie négative), ils sont générés à partir de la distribution log-normale.

Tableau 2. Lois de probabilités et leurs applications possibles

3 Estimation des paramètres

La troisième étape consiste à estimer les paramètres de la distribution sélectionnée. L'estimateur du maximum de vraisemblance (Maximum likelihood estimation : MLE) est le choix préféré pour l'estimation des paramètres, mais d'autres méthodes peuvent être utilisées lorsque le MLE n'a pas une forme simple. Par exemple, le MLE est utilisé pour les distributions exponentielles, normales et log-normales ; la méthode des moments (Method of moment) pour les distributions Erlang, bêta et Weibull.

Distributions	**Méthode d'estimation des paramètres**	**Estimateurs**
$\boldsymbol{Exponentielle\ (\theta)}$	Méthode du maximum de vraisemblance	$\hat{\lambda} = \frac{1}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$
$\boldsymbol{Erlang\ (k,\theta)}$ $\boldsymbol{Gamma\ (k,\theta)}$	Méthode des moments, Méthode du maximum de vraisemblance	$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},\ m_2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2$ $\hat{\lambda} = \frac{1}{\hat{\theta}} = \frac{m_1}{(m_2 - {m_1}^2)}$ $\hat{k} \cong \frac{{m_1}^2}{(m_2 - {m_1}^2)}$
$\boldsymbol{Weibull(\alpha\ ,\ \beta)}$	Méthode des moments, Méthode du maximum de vraisemblance	
$\boldsymbol{Uniforme(a,b)}$		$a = min(x_1, x_2, \dots, x_n)$ $b = max(x_1, x_2, \dots, x_n)$
$\boldsymbol{Bêta\ (\alpha\ ,\ \beta)}$	Méthode des moments	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $\hat{\alpha} = \hat{x}\left(\frac{\hat{x}(1-\hat{x})}{s^2} - 1\right)$ $\hat{\beta} = (1-\hat{x})\left(\frac{\hat{x}(1-\hat{x})}{s^2} - 1\right)$
$\boldsymbol{Normale\ (\mu\ ,\ \sigma)}$ $\boldsymbol{Lognormal\ (\mu\ ,\ \sigma)}$	Méthode du maximum de vraisemblance	$\hat{\mu} = \sum_{i=1}^{n} x_i = \bar{x}$ $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Tableau 3. Résumé de l'ajustement de certaines distributions théoriques

3.1 Fonction `fitdist()`

Le package du langage R **`fitdistrplus`**[1] définit la fonction principale **`fitdist()`**[2] qui permet d'ajuster les parametres d'une distribution théorique à des données en utilisant différentes méthodes d'estimation, notamment la méthode du maximum de vraisemblance (MLE), la méthode des moments, et d'autres méthodes.

Les principaux parametres de la fonction : **`fitdist(data, distr, method ="…")`** :

- **`data`** : l'ensemble de données que vous souhaitez ajuster à une distribution

[1] https://www.rdocumentation.org/packages/fitdistrplus/versions/1.1-11
[2] https://www.rdocumentation.org/packages/fitdistrplus/versions/1.1-11/topics/fitdist

- **`method`**
- **`distr`**: Une chaîne de caractères indiquant le nom de la distribution à ajuster :
  `"exp"`: Distribution exponentielle
  `"beta"` : Distribution beta
  `"weibull"` : Distribution weibull
  `"norm"` : Distribution normale
  `"lnorm"` : Distribution lognormal
- **`method`**: La méthode d'estimation des paramètres de la distribution. Par défaut, la méthode est "mle" (Maximum Likelihood Estimation). D'autres options peuvent inclure "mme" (Method of Moments Estimation) ou d'autres méthodes.

Remarque :

- Si les temps de service ont une plage finie $[a, b]$, la distribution ***Beta*** peut être le choix pour les générer.
- Pour ajuster les parametres de la distribution ***Beta*** il faut transformer vos données pour dans l'intervalle (0, 1), car la distribution bêta est définie sur la plage [0, 1]
- Une approche courante Pour normaliser les valeurs d'un ensemble de données entre 0 et 1 :

$$y_i = (x_i - min(X)) / (max(X) - min(X))$$

où :

- y_i**:** la i^{ème} valeur normalisée dans l'ensemble de données
- x_i **:** la i^{ème} valeur de l'ensemble de données
- $min(X)$: La valeur minimale dans l'ensemble de données
- $max(X)$ **:** La valeur maximale dans l'ensemble de données

```
#Code R : Fonction fitdist
```

```
# Distribution Exponentielle
fExp <- fitdist(data, "exp" , method = "mle")
fExp
```

```
# Distribution Gamma/Erlang
fgamma <- fitdist(data, "gamma" , method = "mle")
fgamma
```

```
# Distribution Beta
dataB<-(data-min(data))/(max(data)-min(data)) #Transformation des valeurs entre [0-1] pour ajuster  une distribution bêta
fbeta <- fitdist(dataB, "beta" , method = "mme")
fbeta
```

```
# Distribution Weibull
fweibull <- fitdist(data, "weibull" , method = "mle")
fweibull
```

```
# Distribution Normale
fnormal <- fitdist(data, "norm" , method = "mle")
fnormal
```

```
# Distribution Lognormale
flnormal <- fitdist(data, "lnorm" , method = "mle")
flnormal
```

4 Test Adéquation de l'ajustement

La quatrième et dernière étape de l'ajustement théorique évalue l'adéquation du modèl.

4.1 Fonction plot

Le tracé d'un objet de classe **fitdist** fournit quatre graphiques classiques d'adéquation :

- **Empirical an theoritical dens. :** Un graphique de densité représentant la fonction de densité de la distribution ajustée ainsi que l'histogramme de la distribution empirique,
- **Empirical an theoritical CDF. :** Un graphique de la fonction de répartition de la distribution empirique et de la distribution ajustée,
- **Q-Q plot :** Un graphique quantile-quantile représentant les quantiles empiriques (axe des y) par rapport aux quantiles théoriques (axe des x), qui permet de comparer la distribution d'un échantillon de données à une distribution théorique
- **P-P plot :** un graphique PP représentant la fonction de distribution empirique évaluée à chaque point de données (axe des y) par rapport à la fonction de distribution ajustée (axe des x).

Une **superposition** étroite entre les graphiques de la distribution empirique (histogrammes/nuages de points) et les courbes de la distribution théorique ajustée **suggère** que la distribution théorique choisie décrit bien la distribution observée.

```
Code R : Fonction plot
```

```
# Distribution Exponentielle
x11()
plot(fExp)
title(main = "Exponontielle", col.main = "red")
```

```
# Distribution Gamma/Erlang
x11()
plot(fgamma)
title(main = "Gamma", col.main = "red")
```

```
# Distribution Beta
x11()
plot(fbeta)
title(main = "Beta", col.main = "red")
```

```
# Distribution Weibull

x11()
plot(fweibull)
title(main = "Weibull", col.main = "red")
```

```
# Distribution Normale
x11()
plot(fnormal)
title(main = "Normal", col.main = "red")
```

```
# Distribution logNormale
x11()
plot(flnormal)
title(main = "Lognormaleormal", col.main = "red")
```

Exemple :

Dans l'exemple de la Figure 4, la distribution exponentielle (à droite) est préférée à la distribution normale (à gauche). En effet, la superposition de la distribution empirique est plus étroite avec la distribution exponentielle qu'avec la distribution normale.

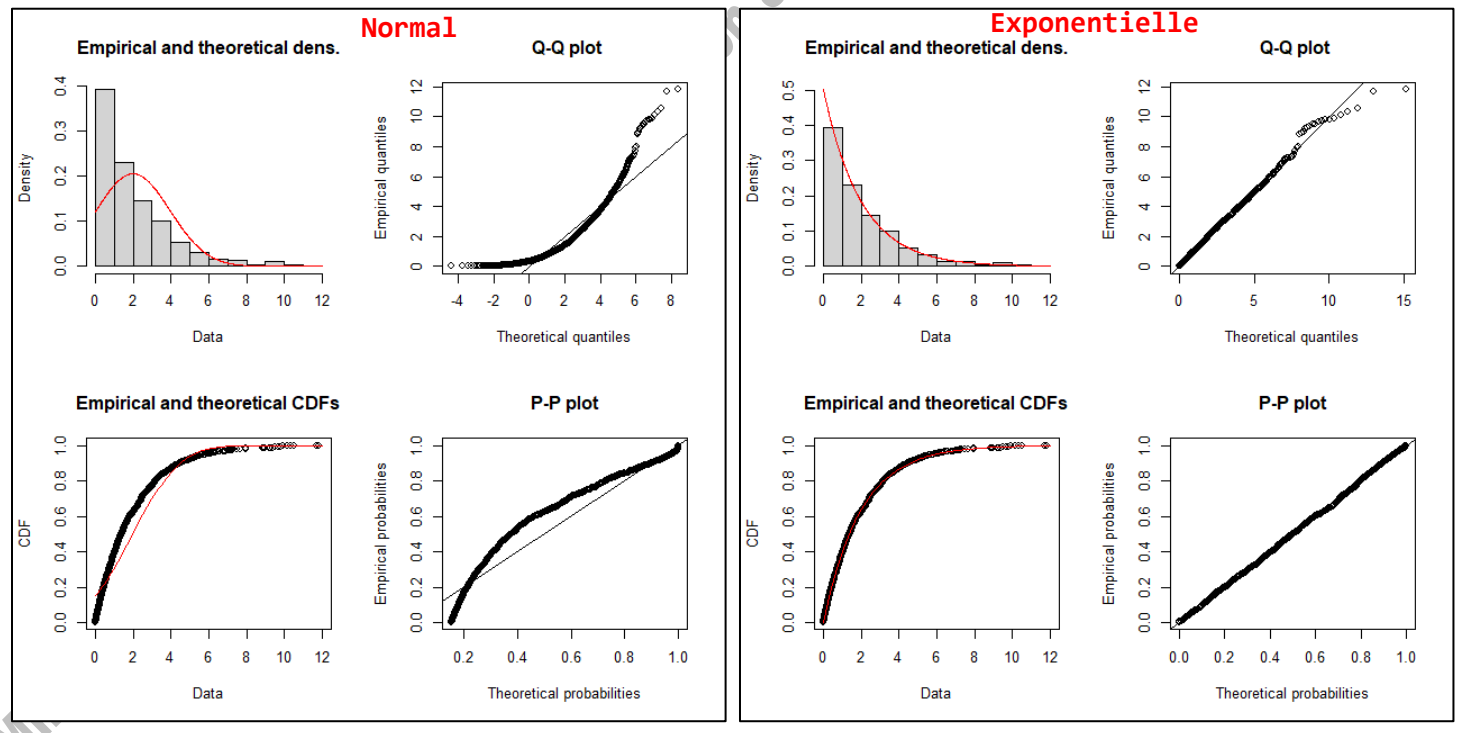


Figure 4. Quatre graphiques d'adéquation pour 2 distributions ajustées (distributions de exponentielle et normale) telles que générées par la fonction **plots()**

4.2 Fonction gofstat

La fonction **`gofstat`** (Goodness-of-Fit Statistics) du package **fitdistrplus** est utilisée pour calculer diverses statistiques d'ajustement (goodness-of-fit) afin d'évaluer la qualité d'un ajustement d'une distribution de théorique aux données observées.

Voici une description des principaux paramètres de la fonction **`gofstat`**:

`gofstat(liste (object1, object2, …), fitnames c(…))`

- **object**: L'objet résultant de la fonction **fitdist**. Il s'agit du résultat de l'ajustement d'une distribution de probabilité aux données.

- **fitnames**: Un vecteur de noms des ajustements. Si spécifié, seuls les ajustements correspondant à ces noms seront inclus dans les statistiques d'ajustement. Si NULL, tous les ajustements dans l'objet seront inclus.

Les tests d'adéquation inclus dans la sortie de **`gofstat()`** sont : ***Kolmogorov-Smirnov*** **(KS),** ***Cramer-von Mises*** **(CvM),** ***Anderson-Darling*** **(AD)**. Ces 3 tests appliquent des méthodes différentes pour mesurer la distance entre la fonction de distribution empirique (basée sur les données observées) et la fonction de distribution théorique ajustée.

Des valeurs plus faibles des statistiques KS, CvM , et AD indique une meilleure adéquation d'un ajustement d'une distribution de théorique aux données observées.

La fonction gofstat() renvoie également des mesures qualité de l'ajustement n à savoir Akaike (AIC) et bayésien (BIC). Ces mesures sont utilisées pour évaluer la qualité de l'ajustement d'une distribution de probabilité aux données observées.

un BIC et un AIC plus faible indique une meilleure qualité d'ajustement.

```
# Code R : Fonction gofstat

gofstat(list(fExp, fgamma, fweibull, fnormal, flnormal),
        fitnames = c("Exp","gamma","weibull","normal","lnormal"))
```

Remarque. La fonction **gofstat()** est appliquée séparément à **fbeta**, car **fbeta** est obtenus sur des données normalisées dans l'intervalle [0,1] (voir section 3.1), tandis que **gofstat()** s'applique aux ajustements obtenus à partir du même jeu de données

Exemple

Dans l'exemple de la Figure 5, les résultats de la fonction indiquent que les distributions exponentielle, weibull, et gamma sont des ajustements meilleurs que les distributions normale et lognormale, car elles présentent les valeurs les plus faibles pour les critères d'information et les statistiques de test d'ajustement.

Remarque. Si les valeurs des paramètres « shape » des distributions gamma et weibull sont égales à 1, ceci indique une distribution exponentielle

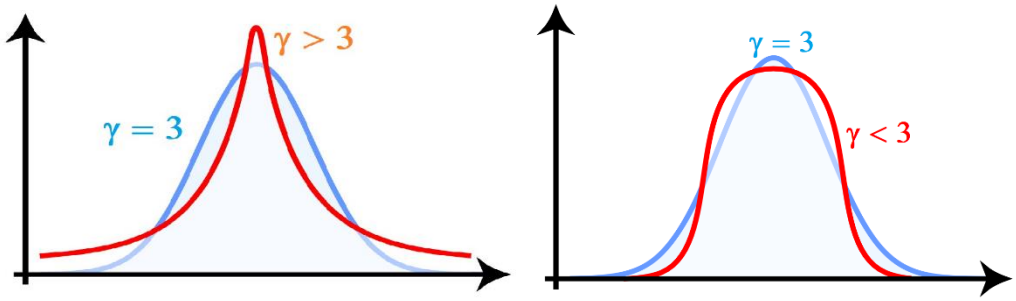
```
Goodness-of-fit statistics
                                    Exp       gamma     weibull      normal     lnormal
Kolmogorov-Smirnov statistic 0.01861793 0.01790487 0.01697985  0.1526966  0.07851413
Cramer-von Mises statistic   0.03797574 0.02759252 0.02539331  7.3154610  2.12011869
Anderson-Darling statistic   0.25861230 0.19192210 0.18217358 43.7817952 12.91114789

Goodness-of-fit criteria
                                    Exp    gamma  weibull   normal  lnormal
Akaike's Information Criterion 3382.072 3383.842 3383.707 4172.552 3592.724
Bayesian Information Criterion 3386.979 3393.657 3393.523 4182.368 3602.540
```

Figure 5. Exemple de résultats renvoyés par la fonction **`gofstat`**

Annexe A: Kurtosis et Skewness

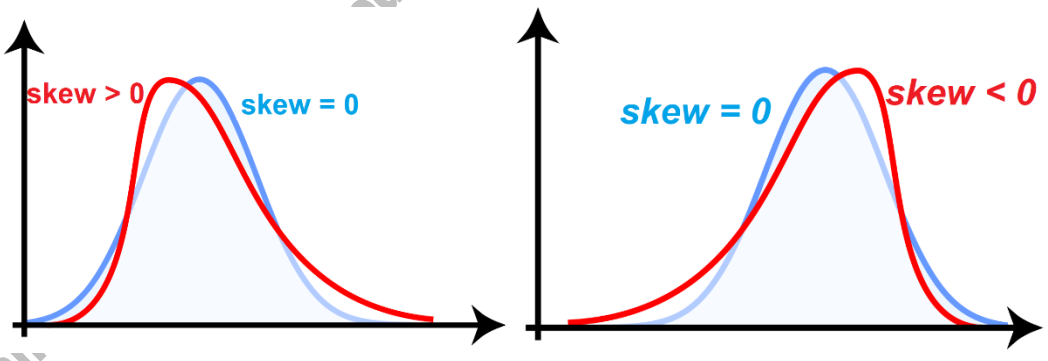
La *Kurtosis* ou coefficient d'aplatissement, noté γ est une mesure statistique qui permet de caractériser l'aplatissement de la distribution.

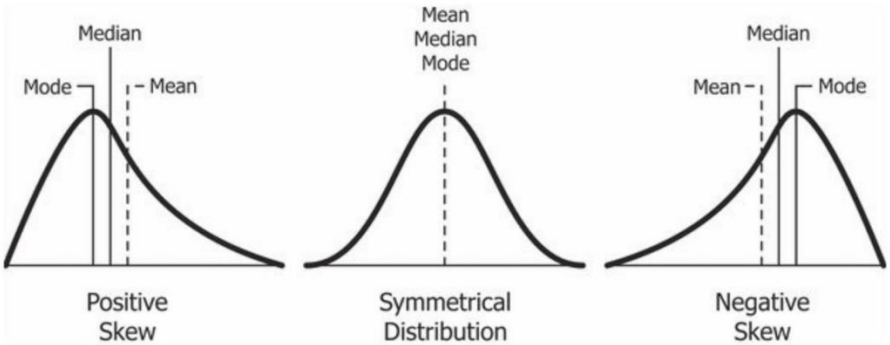
- Une kurtosis positive indique une distribution plus pointue et des queues plus épaisses,
- Une kurtosis négative suggère une distribution plus aplatie.
- Une kurtosis nulle est associée à une distribution normale.



La Skewness mesure l'asymétrie d'une distribution statistique. :

- Une skewness ($skew = 0$)nulle signifie que la distribution est symétrique.
- Une skewness positive ($skew > 0$)indique une queue droite étendue, avec des valeurs extrêmes plus fréquentes à droite de la moyenne.
- Une skewness négative ($skew < 0$) indique une queue gauche étendue, avec des valeurs extrêmes plus fréquentes à gauche.





Annexe B : Fonctions de densité et de répartition des distributions théoriques

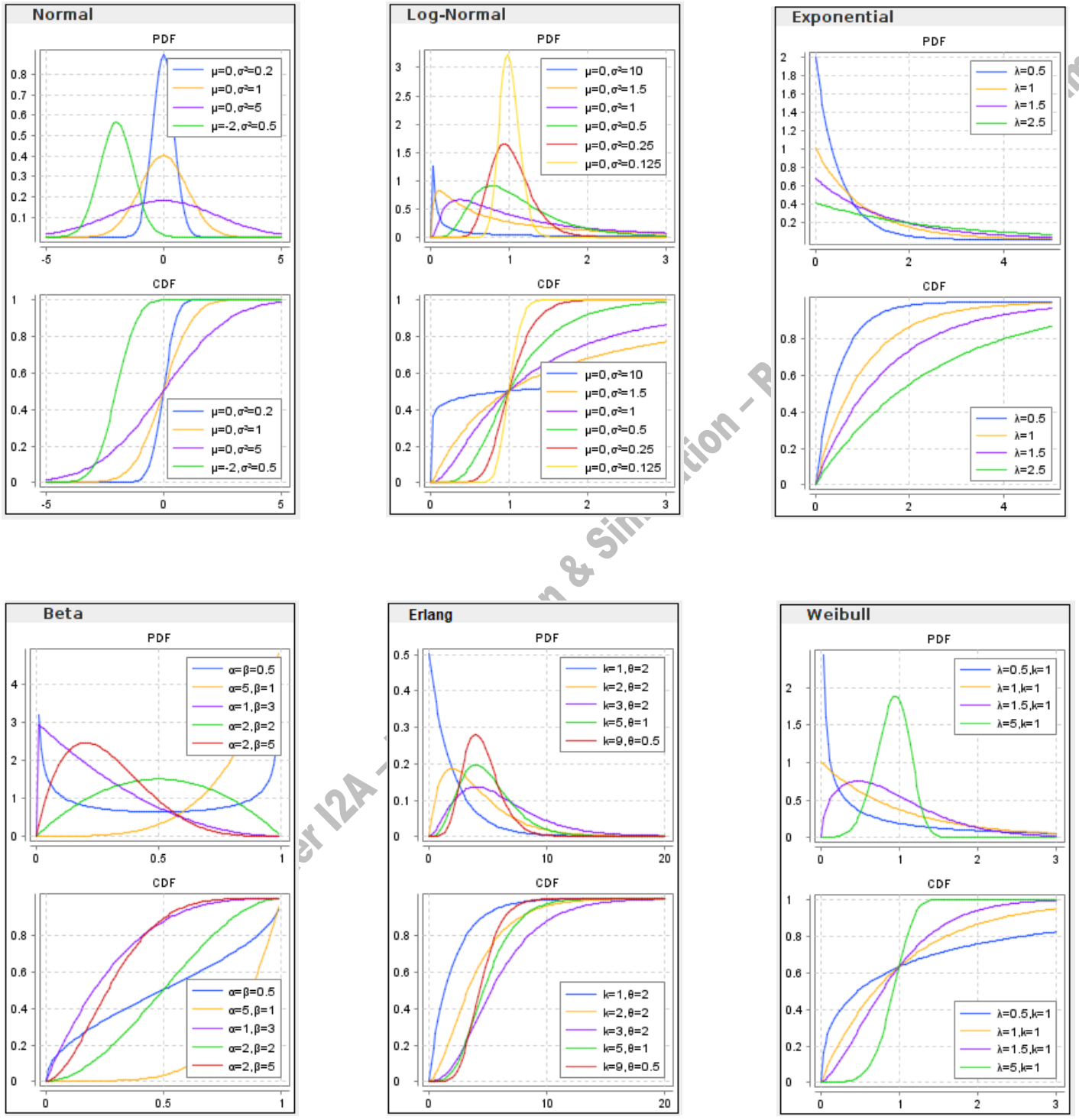


Figure 6. Fonctions de densités (PDF) et fonctions de répartition (CDF) des lois de probabilités continues