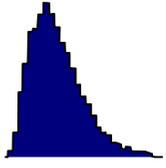


# Spécification d'une loi de probabilité



## Contexte : AQR

### Appréciation quantitative des risques

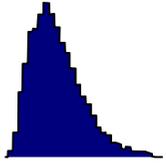
- ~~Approche déterministe~~

~~chaque variable caractérisée par une valeur~~

- Approche stochastique

chaque variable caractérisée par une loi de probabilité

**→ prise en compte des sources de variabilité et/ou d'incertitude**



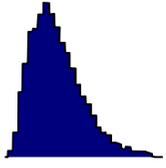
## Comment spécifier chaque loi ?

En fonction

- du type de variable
- des données disponibles

### → **Objectif pédagogique:**

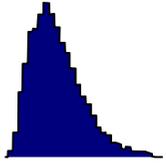
vous présenter une panoplie de lois utiles en AQR avec leur cadre d'utilisation et une méthode pour estimer leurs paramètres



## Outils utilisés

- Logiciel R
  - 2 packages R du projet  
« Risk Assessment with R »  
R. Pouillot, J.B. Denis, M.L. Delignette-Muller
    - **fitdistrplus**  
ajustement de distribution
    - **mc2d**  
Monte Carlo à une ou deux dimensions
- chargeables sur le site du CRAN

<http://lib.stat.cmu.edu/R/CRAN/>



## Définition d'une distribution dans R

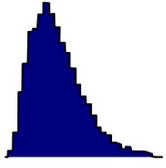
Par les fonctions  $d$ ,  $p$ ,  $q$ ,  $r$  (ex. pour la loi normale)

- **$d$**  : fonction de densité de probabilité  $f$   
ex.: `dnorm(x, mean = 0, sd = 1)`
- **$p$**  : fonction de répartition  $F$  (fréquence cumulée)  
ex.: `pnorm(q, mean = 0, sd = 1)`
- **$q$**  : quantile (associé à la fréquence cumulée  $p$ )  
ex.: `qnorm(p, mean = 0, sd = 1)`
- **$r$**  : échantillonnage aléatoire dans la loi  
ex.: `rnorm(n, mean = 0, sd = 1)`

**Testez ces fonctions avec R**

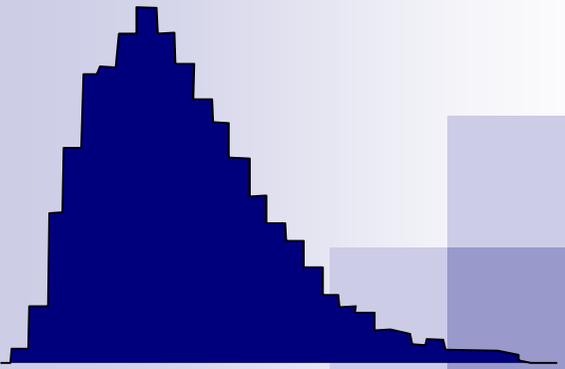
Utilisez les fonctions d,p,q,r avec la loi de votre choix (ici ex. avec loi normale)

```
# fonction dnorm
dnorm(0, mean = 0, sd = 1)
x <- seq(-3, 3, 0.1)
plot(x, dnorm(x), type="l")
# fonction pnorm
pnorm(0, mean = 0, sd = 1)
plot(x, pnorm(x), type="l")
# fonction qnorm
qnorm(0.5, mean = 0, sd = 1)
qnorm(0.975, mean = 0, sd = 1)
p <- seq(0, 1, 0.01)
plot(p, qnorm(p), type="l")
# fonction rnorm
r <- rnorm(1000)
hist(r)
```



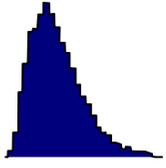
## 4 situations envisagées

1. La théorie probabiliste nous donne la forme et les paramètres de la loi
2. On dispose d'un grand nombre de données observées
3. On doit se baser sur des dires d'experts pour spécifier une loi
4. On veut décrire la loi d'incertitude sur les paramètres d'une loi de variabilité



# 1. Les processus stochastiques

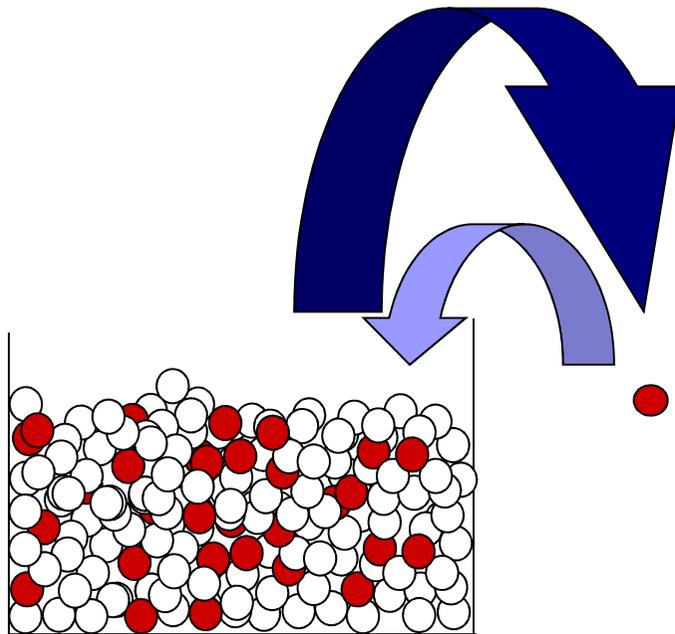
La théorie probabiliste nous donne la forme et les paramètres de la loi



# Le processus binomial

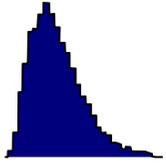
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

**Tirage avec remise**



- $R$  : variable nombre de succès ou réussite  
(ex.: succès si la boule est rouge)
- $n$  : effectif de l'échantillon  
(nb de tirages)
- $p$  : probabilité de succès  
(ici proportion de boules rouges)

$$R \sim \text{Binom}(n, p)$$



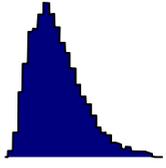
## Utilisation de la loi binomiale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

Nombre de réalisations d'un évènement pour  $n$  tirages successifs indépendants avec une probabilité de succès identique à chaque tirage

Ex.:

- Tirages à pile ou face
- Tirages dans une population d'animaux dont une proportion  $p$  sont infectés si la **population est « très grande au regard de l'échantillon »**

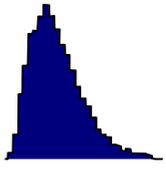


## Description de la loi binomiale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Ecriture:  $R \sim \text{Binom}(n, p)$
- Espérance :  $np$
- Variance :  $np(1 - p)$
- Densité :  $f(r) = C_n^r p^r (1 - p)^{n-r}$

```
rbinom(n=nb_iterations, size=n, prob=p)
```

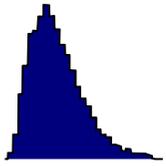


## Estimation du $p$ de la loi binomiale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- par la méthode des moments:

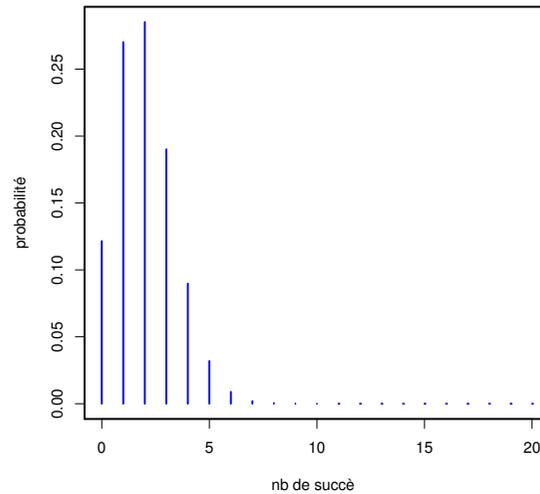
$$\hat{p} = \frac{r}{n}$$



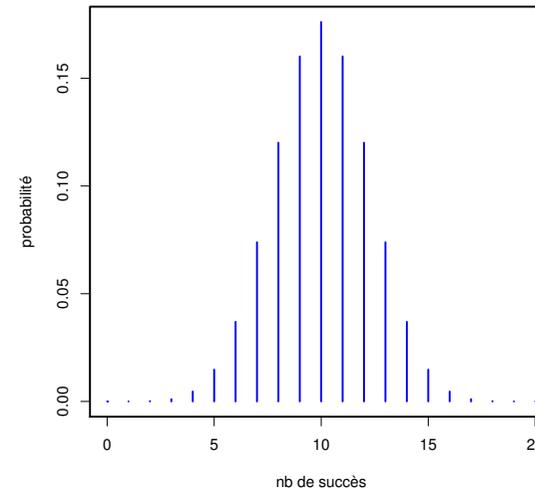
# Visualisation de la loi binomiale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

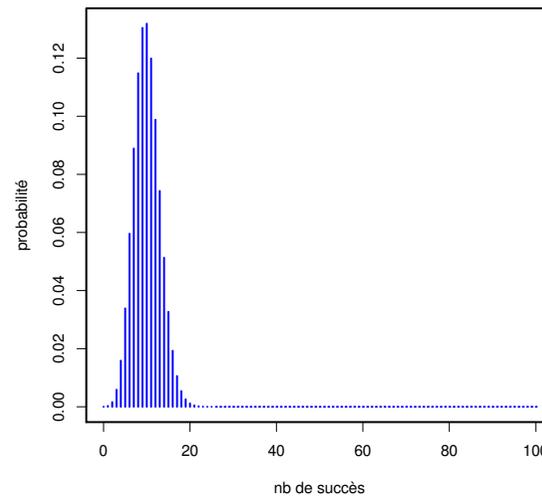
**Binom(20,0.1)**

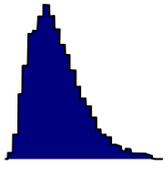


**Binom(20,0.5)**



**Binom(100,0.1)**





## Propriétés asymptotiques de la loi binomiale

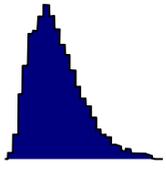
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- La loi binomiale tend vers une **loi normale** lorsque  $np$  et  $n(1-p)$  sont grands (en pratique  $>5$ )

$$\text{Binom}(n, p) \rightarrow N(np, \sqrt{np(1-p)})$$

- La loi binomiale tend vers une **loi de Poisson** lorsque  $n$  est grand et  $np$  est petit (en pratique  $n > 40$  et  $np < 5$ )

$$\text{Binom}(n, p) \rightarrow \text{Poisson}(np)$$



## Ex. loi binomiale 1

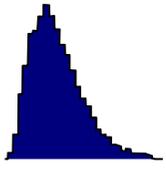
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- La proportion de steaks hachés contaminés par *Escherichia coli* O157:H7 en France est supposée égale à 0.1%.
- On prélève au hasard 500 steaks hachés vendus en France.
- Loi suivie par le nombre de steaks hachés contaminés parmi les prélevés :

loi Binom(500,0.001)

Tracez la densité de probabilité de cette loi avec R

```
plot(0:10, dbinom(0:10, size=500, prob=0.001), type='h')
```



## Ex. loi binomiale 2

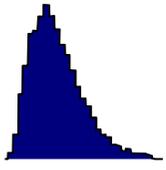
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- On dispose d'un test de détection d'un contaminant donné dans un aliment de sensibilité  $Se = 80\%$
- On teste 500 aliments contaminés.
- Loi suivie par le nombre d'aliments non détectés positifs sur ces 500 aliments contaminés testés :

loi Binom(500,0.2)

Tracez la densité de probabilité de cette loi avec R

```
plot(0:200, dbinom(0:200, size=500, prob=0.2), type='h')
```



# La loi binomiale négative

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

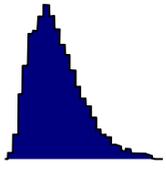
Pour le même processus (probabilité de succès  $p$ ),  
soit  $T$  le nombre de tirages nécessaires à  
l'obtention de  $s$  succès  
( $T-s$  le nombre d'échecs avant  $s$  succès)

- Si le dernier tirage est un succès :  
on arrête de tirer dès qu'on a  $s$  succès

$$T - s \sim \text{NegBin}(s, p)$$

- Si le dernier tirage est de statut inconnu :  
on a eu  $s$  succès et on veut connaître le nombre de tirage  
réalisés

$$T_b - s \sim \text{NegBin}(s+1, p)$$



## Description de la loi binomiale négative

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

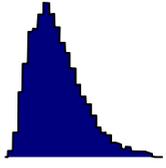
■ Ecriture:  $X \sim \text{NegBin}(s, p)$  avec  $X = T - s$

■ Espérance :  $\frac{s(1-p)}{p}$

■ Variance :  $\frac{s(1-p)}{p^2}$

■ Densité :  $f(x) = C_{s+x-1}^x p^s (1-p)^x$

```
rnbinom(n=nb_iterations, size=s, prob=p)
```



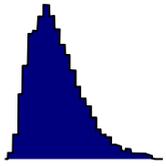
## Estimation du $p$ de la loi binomiale négative

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- par la méthode des moments:

$$\hat{p} = \frac{s}{t}$$

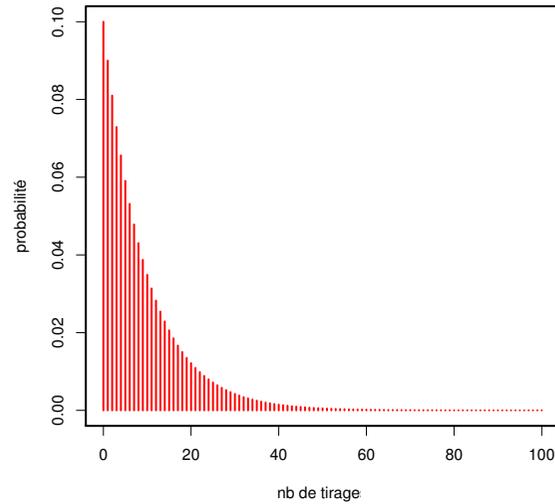
avec  $t$  le nombre de tirages



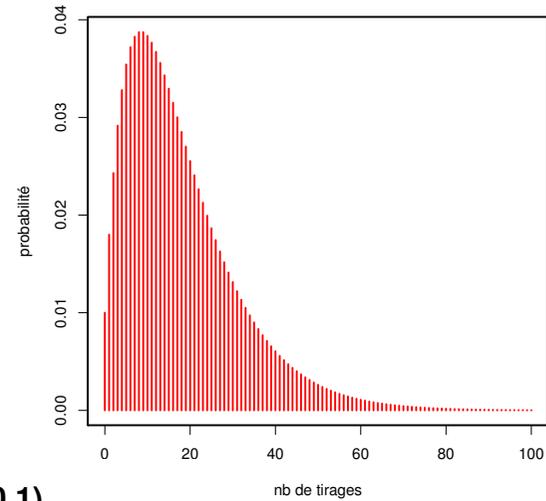
# Visualisation de la loi binomiale négative

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

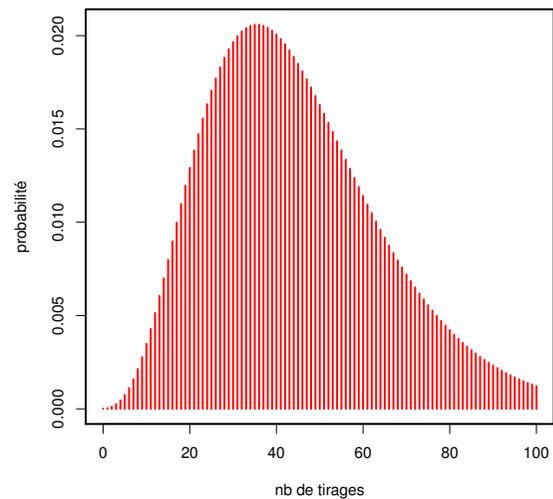
**NegBin(1,0.1)**

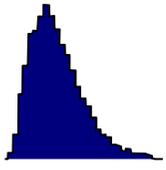


**NegBin(2,0.1)**



**NegBin(5,0.1)**





## Ex. loi négative binomiale 1

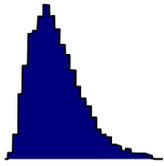
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loix d'incertitude

- Un type d'incident technique donné a une probabilité de 0.1% de se produire lors de chaque fabrication d'un aliment donné.
- Loi suivie par le nombre de fabrications réalisées avant la survenue de ce type d'incident :

$1 + \text{NegBin}(1, 0.001)$ .

Tracez la densité de probabilité de cette loi avec R

```
plot(1+0:2000, dnbinom(0:2000, size=1, prob=0.001), type="h")
```



## Ex. loi négative binomiale 2

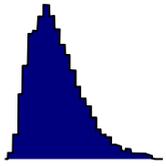
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- On dispose d'un test de détection de contamination par un pathogène donné de lots alimentaires de sensibilité  $Se = 80\%$  et de spécificité 1.
- On teste tous les lots produits et on arrête la production dès qu'un lot est détecté positif.
- Le nombre de lots contaminés produits avant l'arrêt de la production suit la loi :

$$1 + \text{NegBin}(1, 0.80)$$

Tracez la densité de probabilité de cette loi avec R

```
plot(1+0:10, dnbinom(0:10, size=1, prob=0.8), type="h")
```



## Ex. loi négative binomiale 3

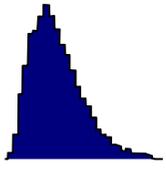
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- On dispose d'un test de détection d'un contaminant donné dans un aliment de sensibilité  $Se = 80\%$  et de spécificité 1.
- Sur 500 aliments de statut inconnu testés par ce test, on détecte 25 aliments positifs.
- Le nombre total d'aliments contaminés parmi les 500 aliments testés suit la loi:

$$25 + \text{NegBin}(25 + 1, 0.80)$$

Tracez la densité de probabilité de cette loi avec R

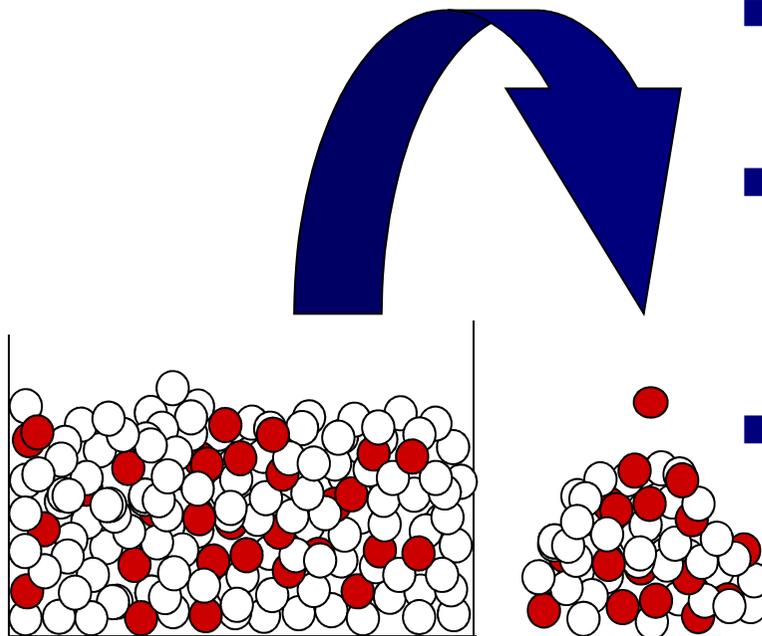
```
plot(25+0:20, dnbinom(0:20, size=25+1, prob=0.8), type="h")
```



# Le processus hypergéométrique

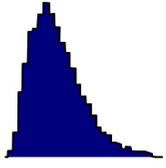
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

**Tirage sans remise**



- $R$  : variable nombre de succès ou réussite  
(ex.: succès si boule rouge)
- $n$  : effectif de l'échantillon  
(nb de tirages)
- $S$  : nombre de succès dans la population  
(ici nb de boules rouges)
- $N$  : effectif de la population

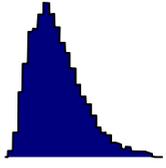
$$R \sim \text{Hypergéom}(N, n, S)$$



## Utilisation de la loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Utilisée pour les tirages sans remise lorsque la taille de l'échantillon dépasse 10% de la population totale
- Si l'échantillon est plus petit, on utilise la loi binomiale



## Description de la loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

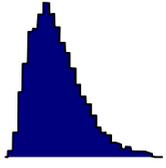
■ Ecriture:  $R \sim \text{Hypergéom}(N, n, S)$

■ Espérance :  $\frac{nS}{N}$

■ Variance :  $\frac{nS(N-S)}{N^2}$

■ Densité :  $f(r) = \frac{C_S^r C_{N-S}^{n-r}}{C_N^n}$

```
rhyper(nn=nb_iterations, m=S, n=N-S, k=n)
```



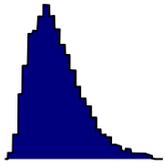
## Estimation du S de la loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Par la méthode des moments

$$\hat{S} = N \frac{r}{n}$$

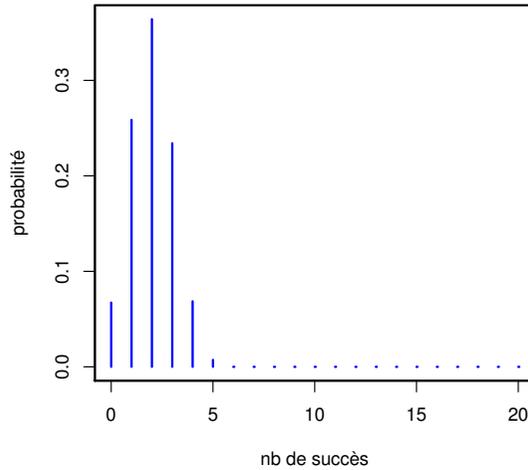
avec  $r$  le nombre de réussites



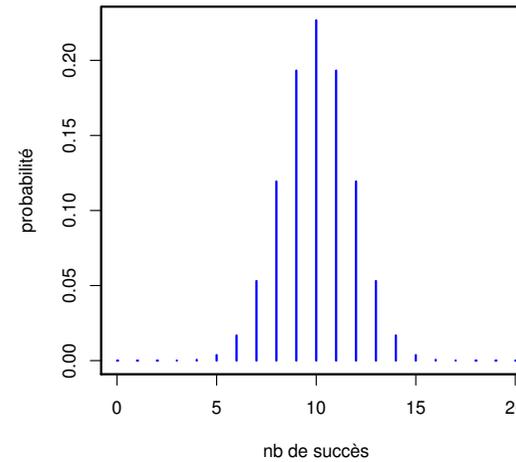
# Visualisation de la loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

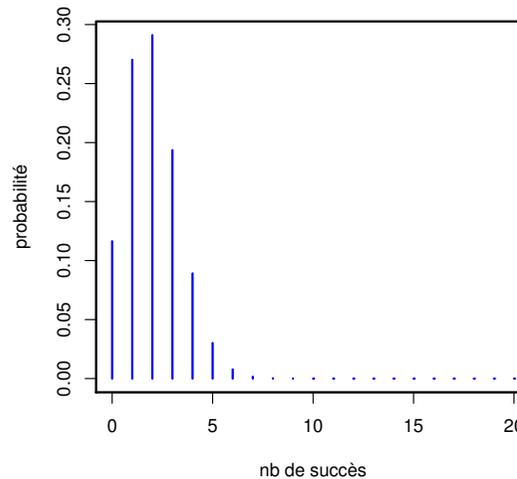
**Hypergéom(50,20,5)**

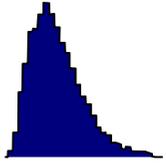


**Hypergéom(50,20,25)**



**Hypergéom(500,20,50)**



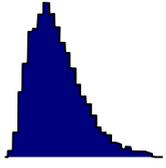


## Propriétés asymptotiques de la loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- La loi hypergéométrique tend vers une **loi binomiale** lorsque  $n$  est petit devant  $N$  (en pratique si  $n/N < 0.1$  )

$$\text{Hypergéom}(N, n, S) \rightarrow \text{Binom}\left(n, \frac{S}{N}\right)$$

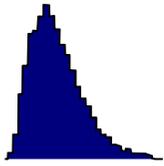


## Ex. loi hypergéométrique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- 50 carcasses de bovins sont utilisées pour la fabrication d'un lot de viande hachée. Parmi ces 50 carcasses, deux sont contaminées par *Escherichia coli* O157:H7.
- On teste la présence du pathogène sur 5 carcasses tirées au hasard parmi les 50.
- Si l'on suppose la sensibilité du test de 100%, le nombre de carcasses détectées comme contaminées suit la loi :

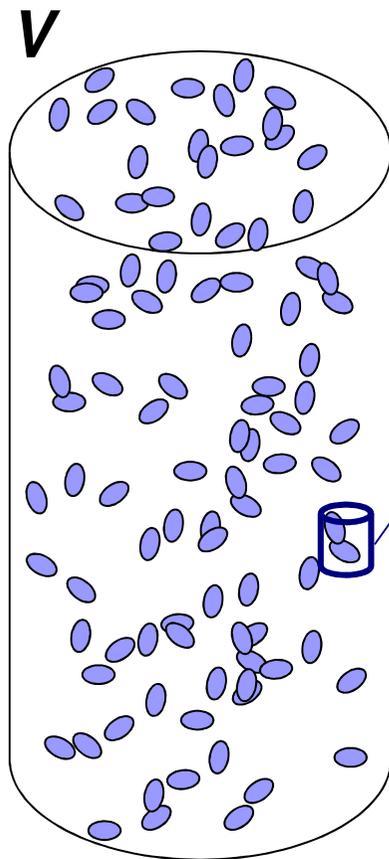
Hypergéom(50,5,2)



# Processus de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

## Répartition aléatoire de $N$ bact.

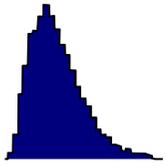


- Ex.: soit une répartition aléatoire homogène de  $N$  bactéries dans un volume  $V$ .
- Soit  $R$  le nb de bactéries dans un petit volume  $v$  pris au hasard dans  $V$ .
- Si  $N$  est grand et  $v$  est petit devant  $V$ ,

$$R \sim \text{Poisson}(\lambda)$$

$$\text{avec } \lambda = N^* v / V$$

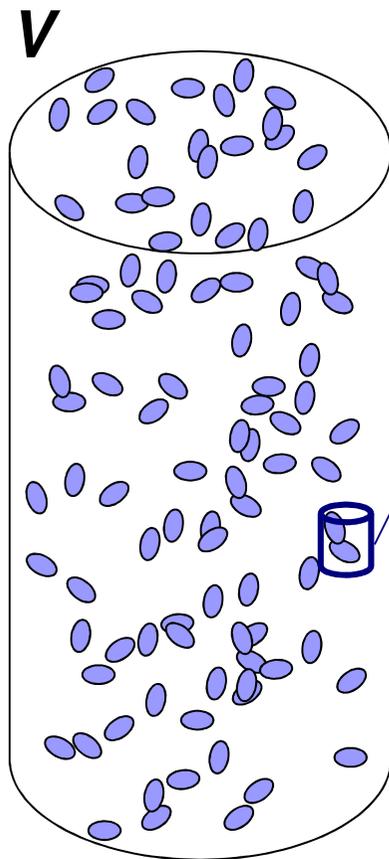
(nb moyen par volume  $v$ )



# De la loi binomiale à la loi de Poisson

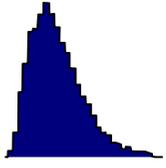
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

## Répartition aléatoire de $N$ bact.



## Démonstration du résultat :

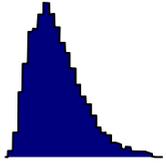
- Chaque cellule a une probabilité  $v/V$  de se trouver dans ce volume
- $R \sim \text{Binom}(N, v/V)$
- Si  $N$  est grand et  $v/V$  est petit  
 $\text{Binom}(N, v/V) \rightarrow \text{Poisson}(\lambda)$   
avec  $\lambda = N * v/V$



## Utilisation de la loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Loi suivie par le nombre d'occurrences d'un évènement dans un intervalle de temps, une surface ou un volume donné,
- caractérisée par sa moyenne (intensité)  $\lambda$
- la probabilité de survenue de l'évènement étant identique dans chaque intervalle de temps, surface ou volume et indépendante de la survenue de l'évènement dans un autre intervalle de temps, surface ou volume



## Description de la loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

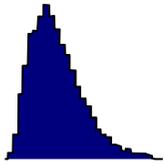
■ Ecriture:  $R \sim \text{Poisson}(\lambda)$

■ Espérance :  $\lambda$

■ Variance :  $\lambda$

■ Densité :  $f(r) = \frac{e^{-\lambda} \lambda^r}{r!}$

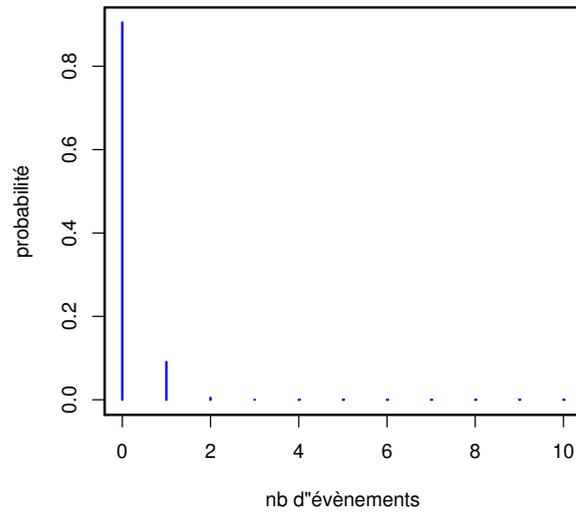
```
rpois(n=nb_observations, lambda= $\lambda$ )
```



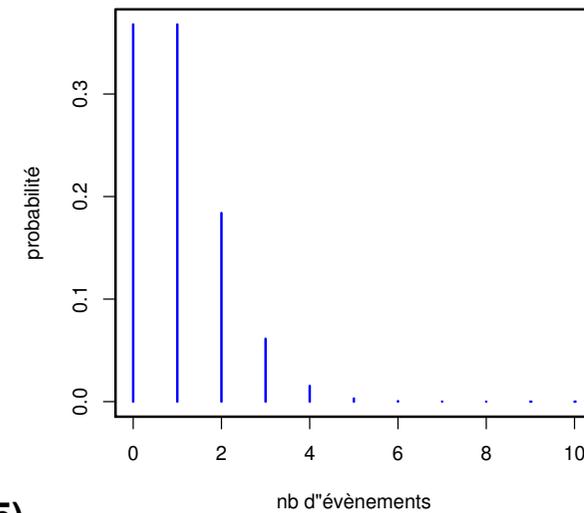
# Visualisation de la loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

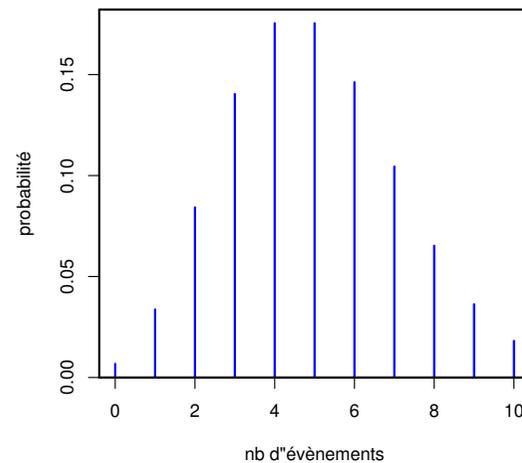
**Poisson(0.1)**

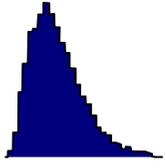


**Poisson(1)**



**Poisson(5)**



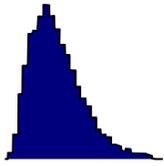


## Propriétés asymptotiques de la loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- La loi de Poisson tend vers une **loi normale** lorsque  $\lambda$  est grand (en pratique  $>5$ )

$$\text{Poisson}(\lambda) \rightarrow N(\lambda, \sqrt{\lambda})$$



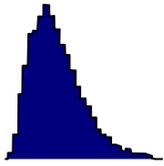
## Ex.1 loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- On mélange 25g d'une matrice alimentaire contaminée à 1 UFC.g<sup>-1</sup> par une bactérie donnée à 225ml de bouillon de culture.
- On prélève 1 ml du mélange obtenu (mélange dans lequel on suppose les bactéries réparties aléatoirement de façon homogène).
- Le nombre de bactéries présentes dans le volume prélevé suit la loi:

Poisson(0.1)

```
plot(0:5, dpois(0:5, lambda = 0.1), type="h")
```



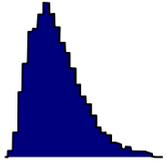
## Ex. 2 Loi de Poisson

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Sur une chaîne de fabrication donnée, un type d'incident technique donné se produit de manière totalement aléatoire (indépendamment des incidents précédents) avec une fréquence moyenne de 10 par jour (8h de fabrication).
- Le nombre d'incidents se produisant en une heure suit la loi :

**Poisson(10/8=1.25)**

```
plot(0:10, dpois(0:10, lambda = 1.25), type="h")
```



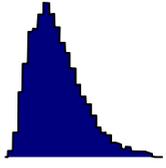
# La loi exponentielle

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Pour le même processus de Poisson, soit  $X$  le volume (ou temps) nécessaire à l'observation d'un évènement ,

$$X \sim \text{Expo}(1/\lambda)$$

avec  $\lambda$  le nombre moyen d'évènements par unité de volume ou temps

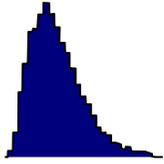


## Description de la loi exponentielle

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Ecriture:  $X \sim \text{Expo}(\beta)$
- Espérance :  $\beta$
- Variance :  $\beta^2$
- Densité :  $f(x) = \frac{\exp(-x/\beta)}{\beta}$
- Fonction de répartition:  $F(x) = 1 - \exp(-x/\beta)$

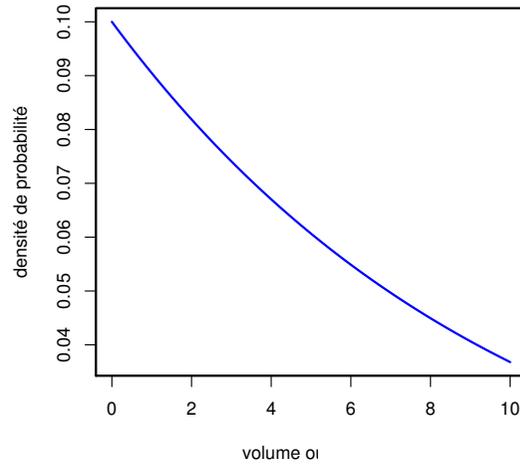
```
rexp(n = nb_observations, rate = 1/β = λ)
```



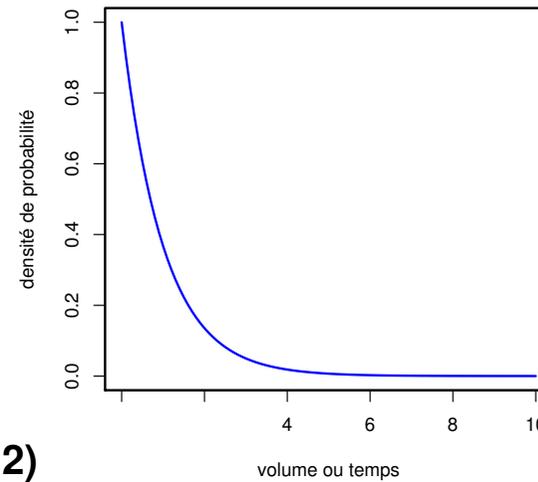
# Visualisation de la loi exponentielle

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

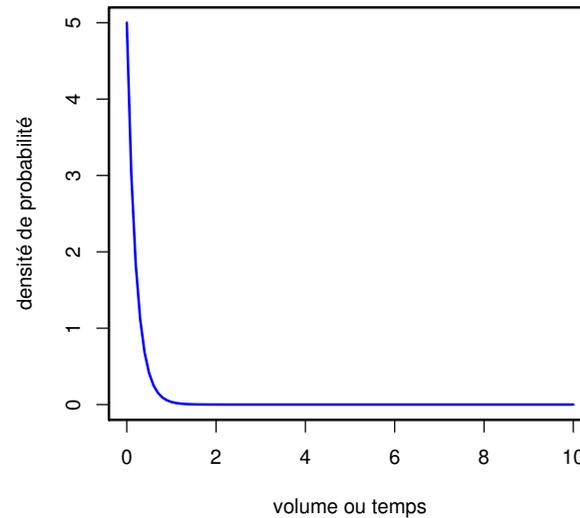
**Expo(10)**

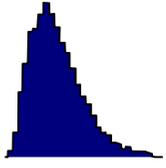


**Expo(1)**



**Expo(0.2)**





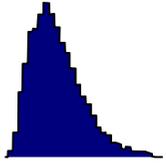
## Ex. loi exponentielle

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Sur une chaîne de fabrication donnée, un type d'incident technique donné se produit de manière totalement aléatoire (indépendamment des incidents précédents) avec une fréquence moyenne de 10 par jour (8h de fabrication).
- Le temps de fabrication avant observation d'un incident suit la loi :

**Expo(0.8)**

```
plot(0:10, dexp(0:10, rate = 0.8), type="l")
```



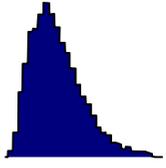
# La loi gamma

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Plus généralement, pour ce même processus de Poisson, soit  $X$  le volume (ou temps) nécessaire à l'observation de  $\alpha$  évènements,

$$X \sim \text{Gamma}(\alpha, \beta = 1/\lambda)$$

avec  $\lambda$  le nombre moyen d'évènements par unité de volume ou temps



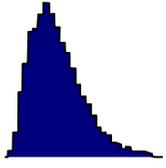
## Description de la loi gamma

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Ecriture:  $X \sim \text{Gamma}(\alpha, \beta)$   
 $\alpha$  paramètre de forme,  $\beta$  paramètre d'échelle  
(tous deux positifs)
- Espérance :  $\alpha\beta$
- Variance :  $\alpha\beta^2$
- Densité :  $f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} \exp(-x/\beta)}{\Gamma(\alpha)}$

avec  $\Gamma$  la fonction Gamma

```
rgamma (n=nb_iterations, shape =  $\alpha$ , rate =  $1/\beta = \lambda$ )
```



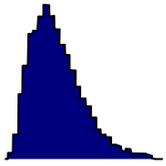
## Estimation des paramètres de la loi gamma

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Par la méthode des moments:

$$\hat{\alpha} = \frac{\bar{x}^2}{\text{Var}(x)}$$

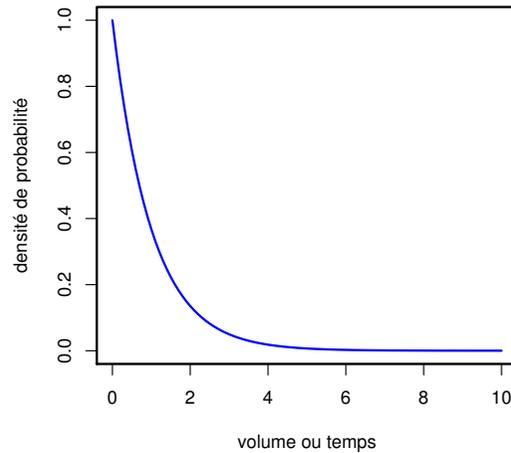
$$\hat{\beta} = \frac{\text{Var}(x)}{\bar{x}}$$



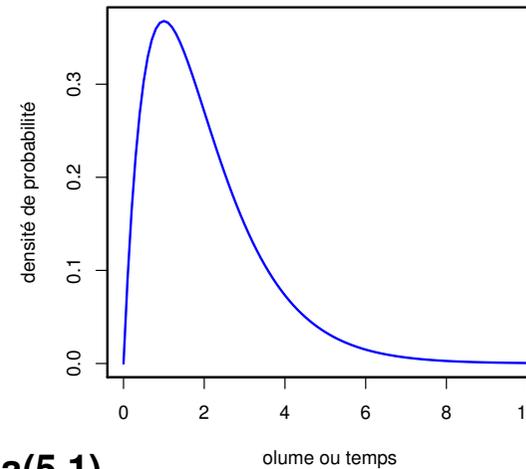
# Visualisation de la loi gamma

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

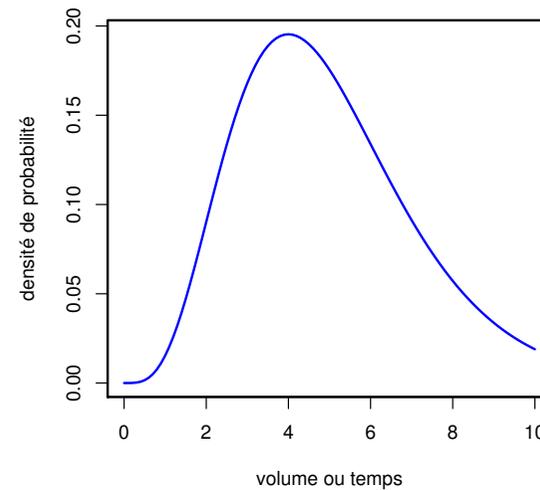
**Gamma(1,1)**

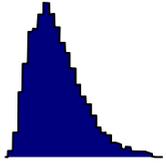


**Gamma(2,1)**



**Gamma(5,1)**

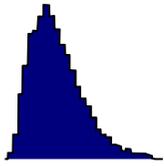




## Revenons sur la loi négative binomiale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Classiquement utilisée pour modéliser la loi du nombre de bactéries par échantillon issus d'un milieu peu homogène  
surdispersion de la loi de Poisson
- **Loi négative binomiale appelée loi gamma-Poisson**  
= mélange des lois Poisson et gamma  
= loi de Poisson caractérisé par  $\lambda$  variant suivant une loi gamma

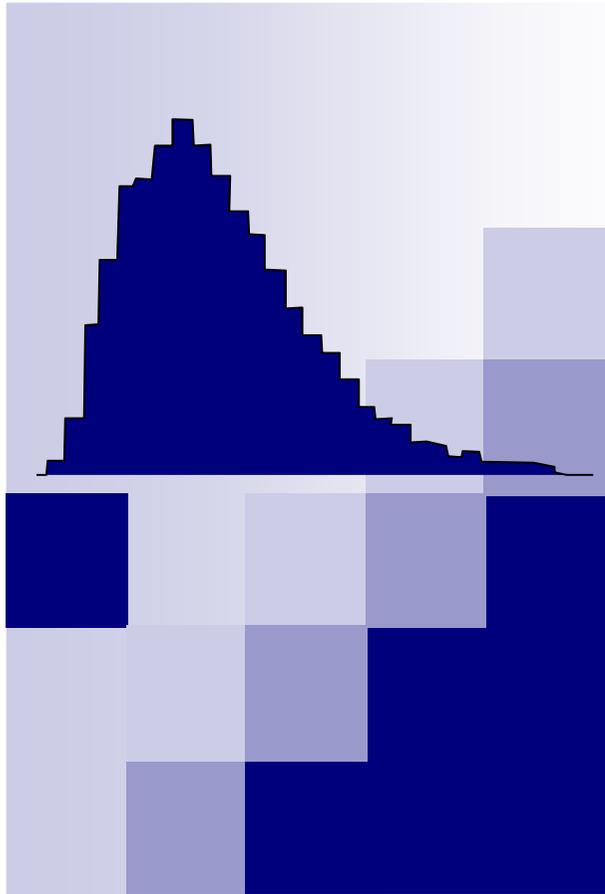


## A retenir



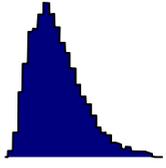
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Tirage avec remise ou dans une grande population
  - Nb. de succès : **loi binomiale**
  - Nb. d'échecs avant s succès : **loi binomiale négative**
- Tirage sans remise
  - Nb de succès : **loi hypergéométrique**
- Répartition homogène de cellules
  - Nb. de cellules dans un volume : **loi de Poisson**
  - Volume nécessaire à l'obtention d'une cellule : **loi exponentielle**
  - de n cellules : **loi gamma**



## 2. On dispose d'un grand nombre de données observées

- utilisation directe des données
- ajustement d'une loi sur les données



## Loi directement décrite par les données

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

Tirage parmi les valeurs observées

- **Avantage :**

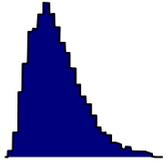
on évite les biais possibles liés à l'ajustement d'une loi non adéquate

- **Inconvénient:**

les données ne reflètent pas l'ensemble des possibilités mais seulement un échantillon  
(occurrence des valeurs extrêmes peu probable)

Adapté aux variables qualitatives ou quantitatives discrètes avec peu de classes ou très nombreuses données

```
sample (x=vecteur_des_observées, size=nb_iterations, replace=TRUE)
```



## Modélisation non paramétrée de la fonction de répartition $F$

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

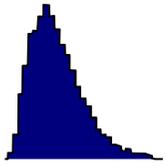
Solution améliorée pour les variables  
quantitatives continues

- Valeurs  $x_{\min}$  et  $x_{\max}$  définies par dires  
d'expert

$$F(x_{\min}) = 0 \quad F(x_{\max}) = 1$$

- Données observées pour définir  $F$  en  
chacune des  $n$  valeurs croissantes  $x_i$

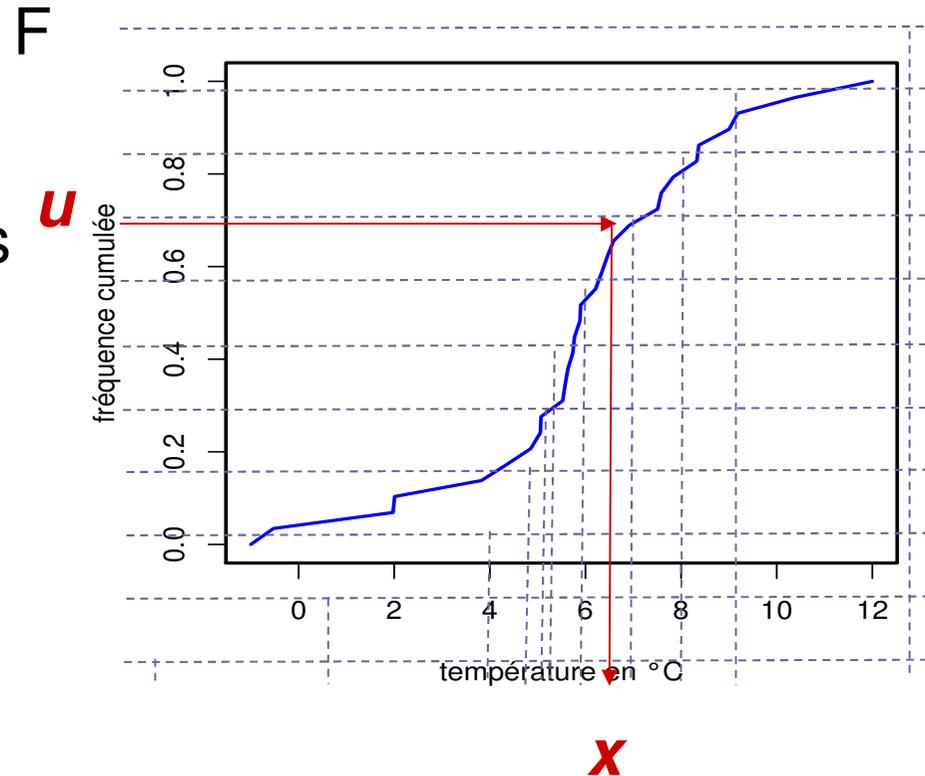
$$F(x_i) = \frac{i}{n+1}$$



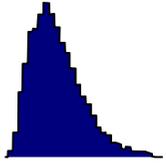
## Puis tirage dans la loi définie par $F$

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loix d'incertitude

- Tirage de  $u$  dans la loi uniforme  $\text{Unif}(0,1)$
- Calcul de  $x$  tel que  $F(u) = x$



```
rempiricalC(n= nb_iterations, min, max, values = vecteur_des_observées)
```

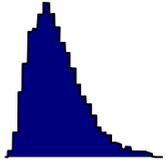


## Loi paramétrée ajustée sur les données

1. Processus stochastiques
- 2. Nombreuses données**
3. Dires d'expert
4. Loïs d'incertitude

### Etapes conseillées

1. Choix de lois paramétrées candidates à l'ajustement par les lois classiques
2. Ajustement de ces lois aux données
3. Sélection de la loi ajustant le mieux les données

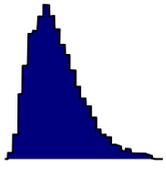


## Lois paramétrées ajustables aux données observées

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

Choisies parmi les lois classiques susceptibles de décrire la variable en fonction

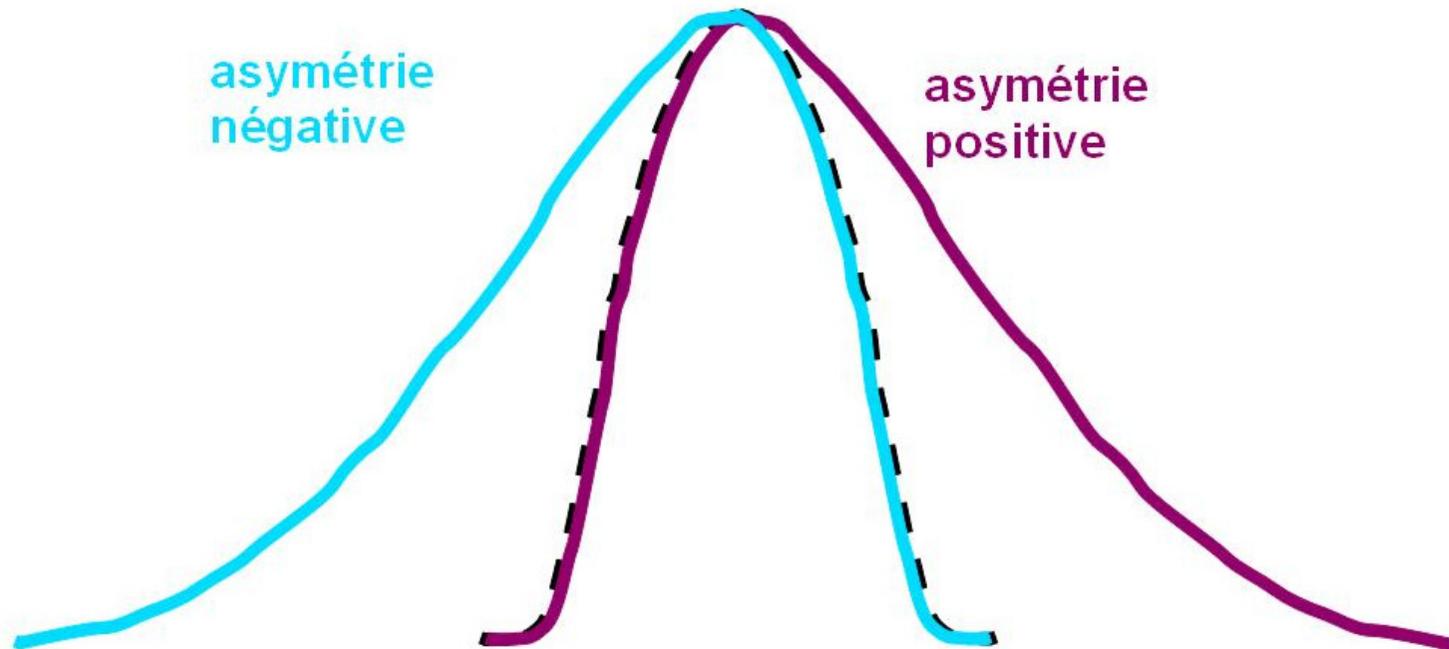
- du caractère continu ou discret de la variable
- de son domaine de définition
- parfois de la connaissance de la loi généralement suivie par ce type de variable
- des coefficients de dissymétrie (skewness) et d'aplatissement (kurtosis) en comparaison de ceux des distributions classiques

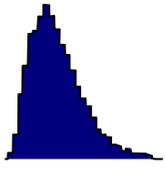


# Skewness (coefficient d'asymétrie)

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

$$skewness = \frac{\frac{1}{n} \times \sum (x_i - \mu)^3}{\sigma^3}$$

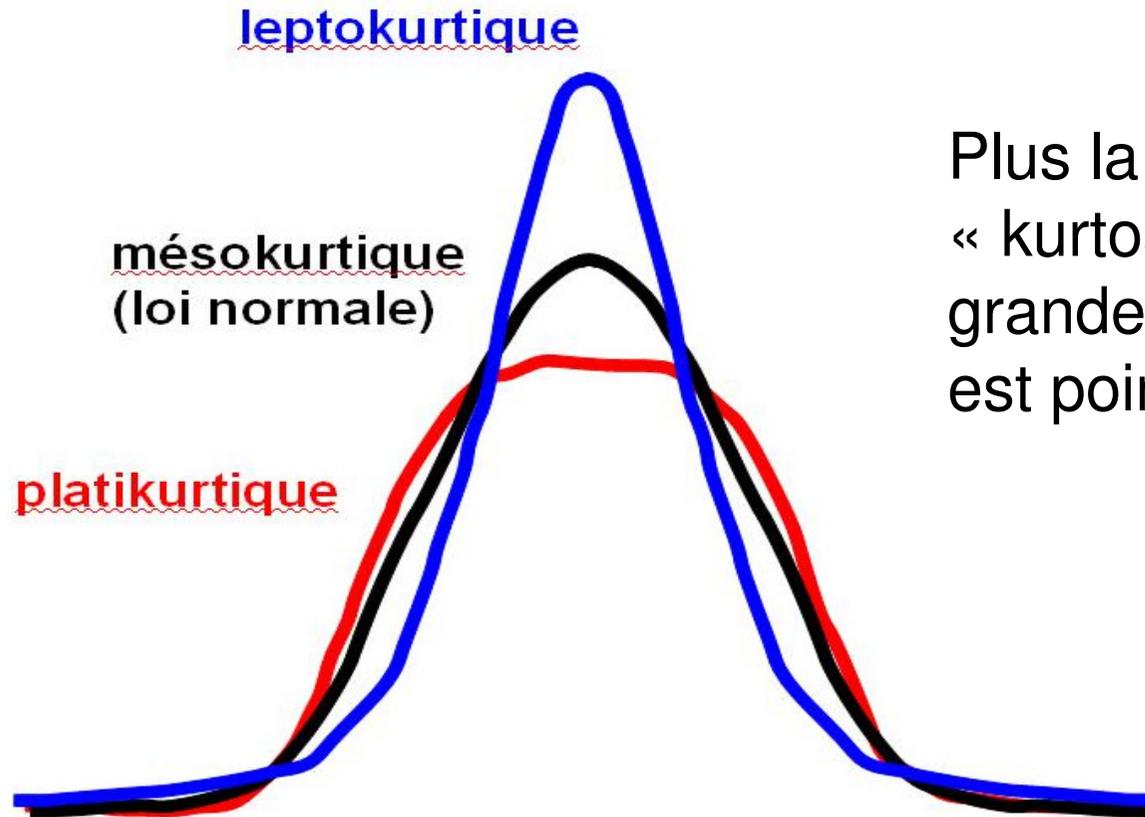




# Kurtosis (pointicité – applatissement)

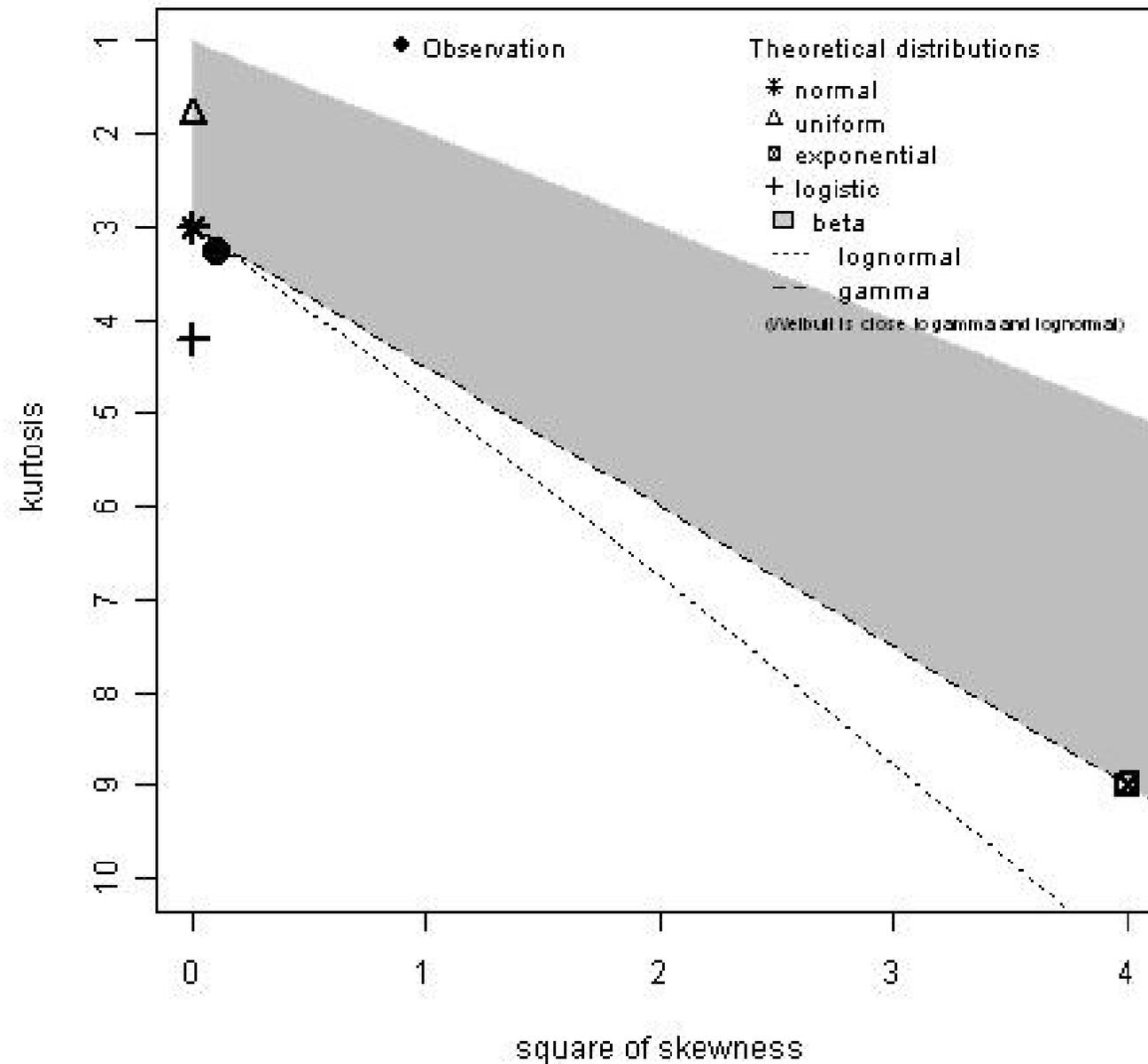
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loix d'incertitude

$$kurtosis = \frac{\frac{1}{n} \times \sum (x_i - \mu)^4}{\sigma^4}$$

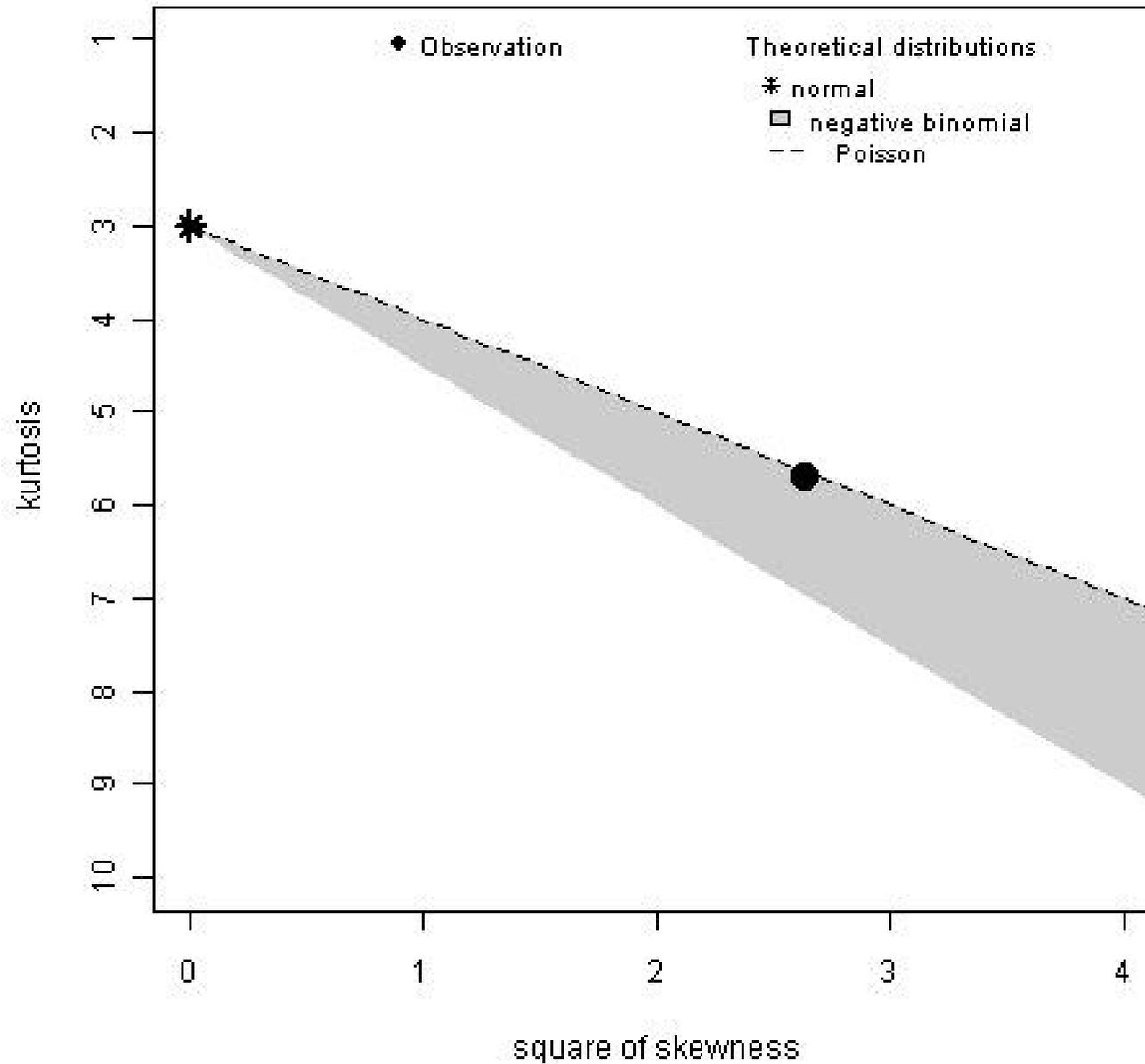


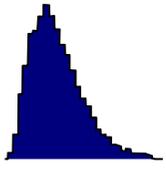
Plus la  
« kurtosis » est  
grande plus la loi  
est pointue

# Ex. de graphe de Cullen et Frey pour loi continue



# Graphe de Cullen et Frey pour loi discrète





# Méthodes d'ajustement

1. Processus stochastiques
2. **Nombreuses données**
3. Dires d'expert
4. Loïs d'incertitude

## ■ Méthode des moments

- Ajustement de moments (moyenne, écart type, indice de dissymétrie ...) de la loi aux moments observés

(autant de moments qu'il y a de paramètres caractérisant la loi)

## ■ Méthode du maximum de vraisemblance

- Recherche des paramètres permettant d'obtenir la plus forte vraisemblance des données observées = la plus forte probabilité de l'ensemble des valeurs observées
- **Utilisable même sur des données censurées**

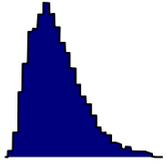
## Maximum de vraisemblance

Pour une distribution de paramètre  $\theta$  on choisit  $\theta$  qui maximise la vraisemblance des données  $x$

$$\max \left( \prod_{i=1}^n \Pr(x_i | \theta) \right)$$

Dans le cas d'une variable quantitative continue

$$\max \left( \prod_{i=1}^n f(x_i | \theta) \right)$$

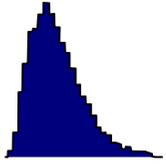


## Choix d'une loi parmi les lois ajustées

1. Processus stochastiques
2. **Nombreuses données**
3. Dires d'expert
4. Lois d'incertitude

Quelle loi décrit-elle le mieux les données ?

- Observation graphique :  
Graphes d'ajustement



# Graphes d'ajustement

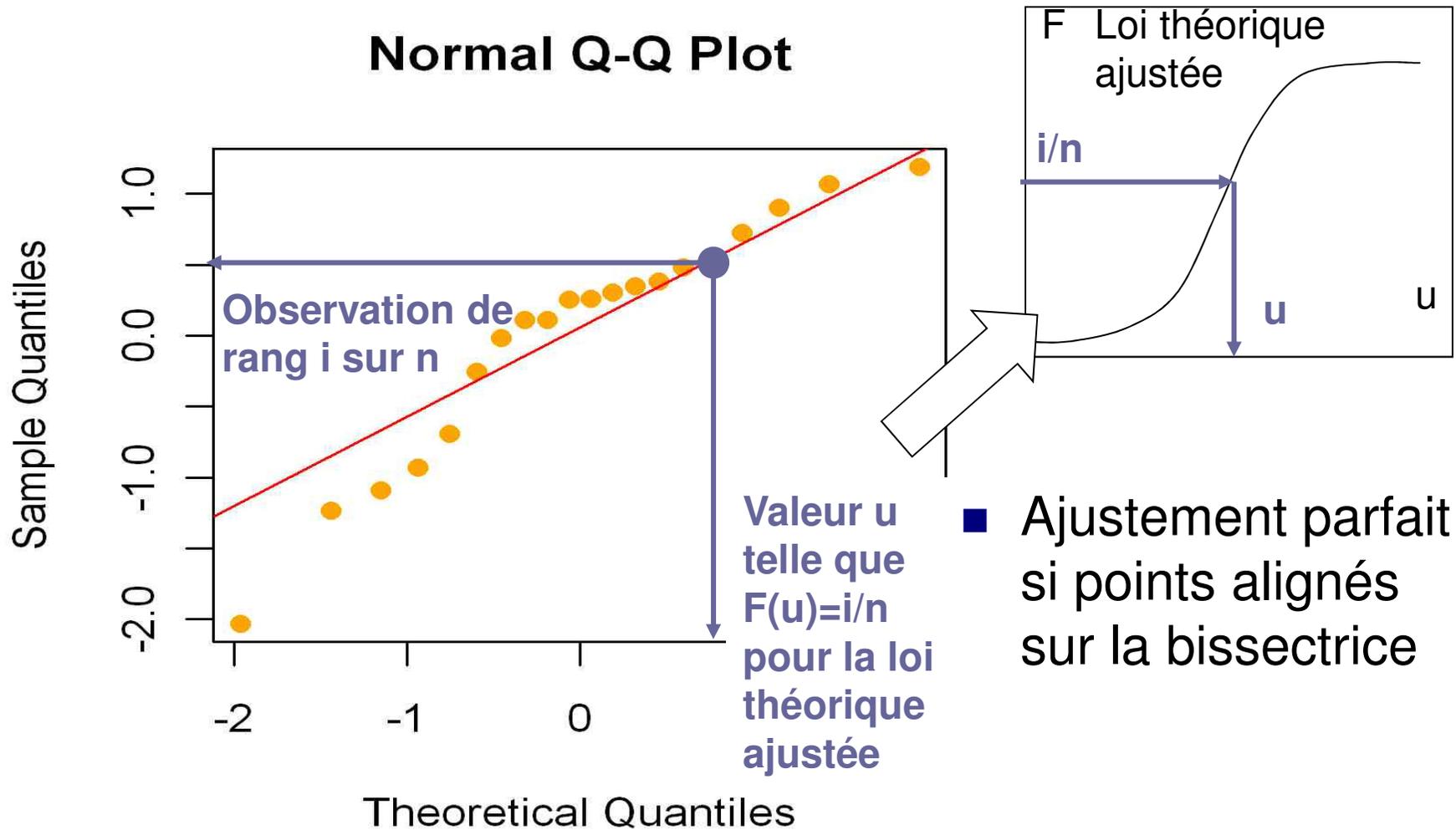
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

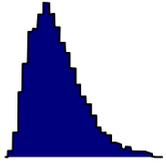
- Fonction de densité de probabilité théorique  $f$  superposée à l'histogramme
- Fonction de répartition théorique  $F$  superposée à la courbe des fréquences cumulées
- Diagramme des fréquences cumulées observées en fonction des théoriques : PP-plot
- Diagramme des quantiles observés en fonction des quantiles théoriques : **QQ-plot**  
souvent préféré au PP-plot car accentuant les écarts entre les deux distributions au niveau de leurs queues

*Toutes les valeurs décrites par la loi sont-elles réalistes?*

*Toutes les valeurs réalistes sont-elles décrites par la loi?*

# Rappel sur le QQ-plot





## Choix d'une loi parmi les lois ajustées

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

Quelle loi décrit-elle le mieux les données ?

- Observation graphique :  
Graphes d'ajustement
- Utilisation de critères d'ajustement
  - Log-vraisemblance, AIC, BIC
  - $\chi^2$  (adapté aux lois discrètes)
  - Kolmogorov-Smirnov
  - Cramer-Von Mises, **Anderson-Darling**

# Statistiques classiques basées sur la vraisemblance

## Log-vraisemblance

$$\text{loglikelihood} = \ln\left(\prod_{i=1}^n f(x_i|\theta)\right) = \sum_{i=1}^n \ln(f(x_i|\theta))$$

**Critères pénalisant la complexité du modèle par le nombre de paramètres (*npar*) de la distribution**

### **Critère d'information d'Akaïké**

$$AIC = -2 \times \text{loglikelihood} + 2 \times npar$$

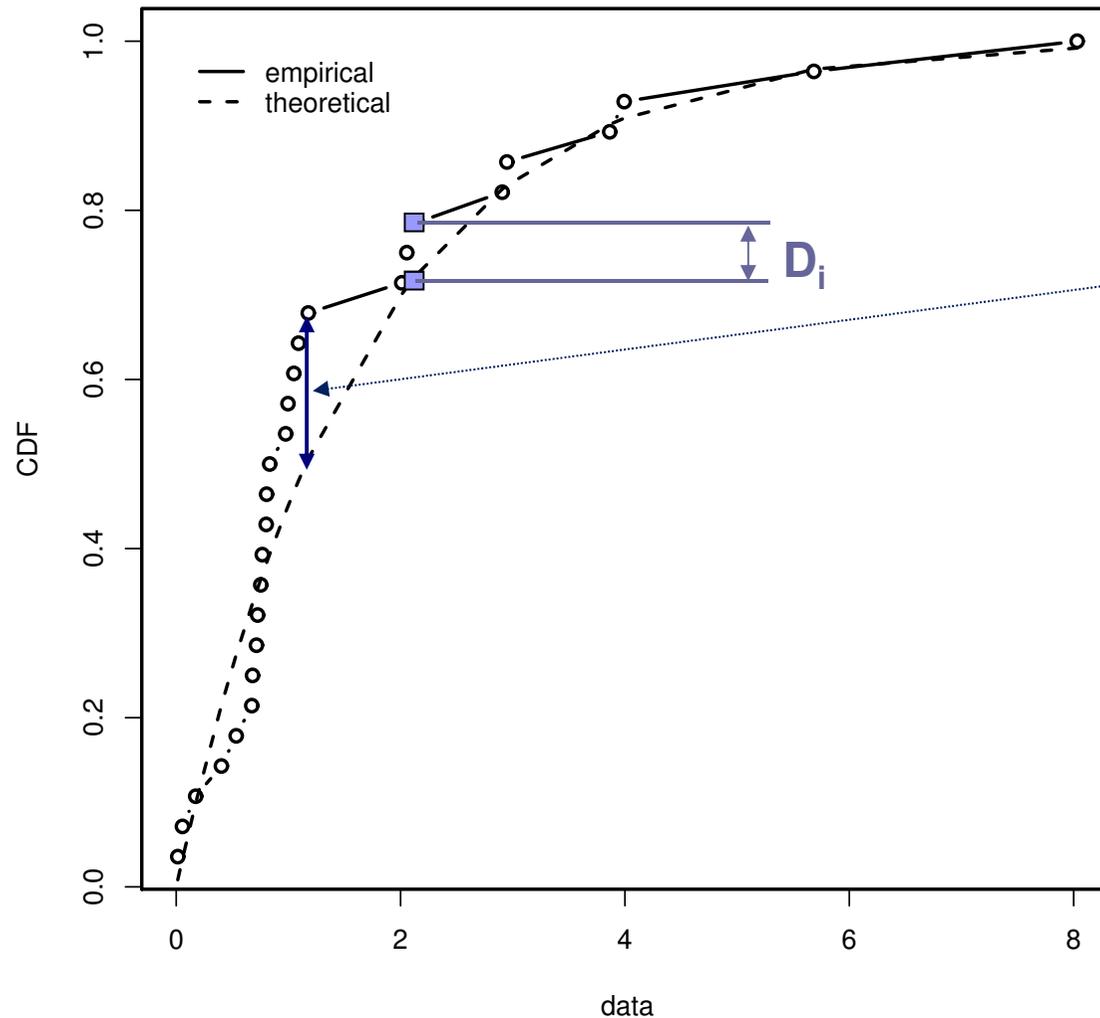
### **Critère bayésien d'information (Schwarz)**

$$BIC = -2 \times \text{loglikelihood} + \ln(n) \times npar$$

# Statistique de Kolmogorov-Smirnov

$$KS = \max(|F_{obs}(x_i) - F_{theo}(x_i)|) = \max(D_i)$$

Empirical and theoretical CDFs



**KS**

Peu de poids accordé  
aux queues de  
distribution  
(écarts généralement  
plus faibles en queues  
de distribution)

## Statistique de Cramer-Von Mises

### Distance quadratique entre les fonctions de répartition

$$\omega^2 = n \int_{-\infty}^{\infty} (F_{obs}(x) - F_{theo}(x))^2 dx$$

Assez rarement utilisée en appréciation des risques

Là encore peu de poids accordé aux écarts sur les  
queues de distribution

## Statistique d'Anderson-Darling

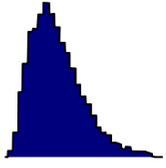
$$A^2 = n \int_{-\infty}^{\infty} (F_{obs}(x) - F_{theo}(x))^2 \psi(x) dx$$

**distance quadratique avec pondération  
donnant du poids aux queues de distribution**

$$\psi(x) = (F_{theo}(x) \times (1 - F_{theo}(x)))^{-1}$$

À préférer en appréciation des risques où les queues de distributions ont généralement un poids important sur le risque

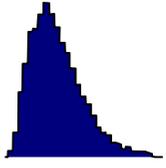
Mais attention, contrairement aux 2 précédentes, **cette distance est dépendante de la loi théorique** (par les poids) donc la comparaison entre les ajustements de plusieurs distributions par comparaison de cette statistique est délicate !



## Quelques lois classiques variable quantitative continue

1. Processus stochastiques
2. **Nombreuses données**
3. Dires d'expert
4. Lois d'incertitude

- Loi normale (définie sur  $[-\infty; +\infty]$ )
- Loi log-normale (définie sur  $[0; +\infty]$ )
- Loi exponentielle (définie sur  $[0; +\infty]$ )
- Loi bêta (définie sur  $[0; 1]$ )
- Loi gamma (définie sur  $[0; +\infty]$ )
- Loi de Weibull (définie sur  $[0; +\infty]$ )
- Loi logistique (définie sur  $[-\infty; +\infty]$ )



## Description de la loi normale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

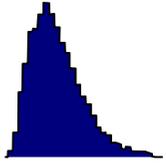
■ Ecriture :  $X \sim N(\mu, \sigma)$

■ Espérance :  $\mu$

■ Variance :  $\sigma^2$

■ Densité :  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

```
rnorm(n=nb_iterations, mean=μ, sd=σ)
```



# Estimation des paramètres de la loi normale

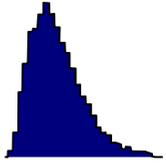
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

par la méthode du maximum de  
vraisemblance:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

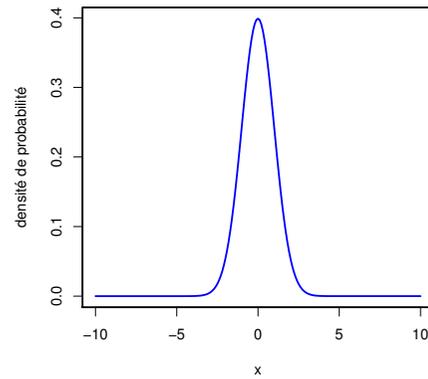
Estimateur  
biaisé de la  
variance



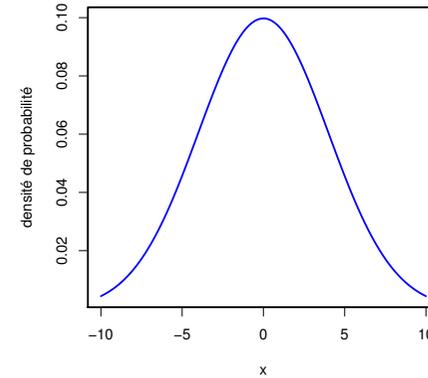
# Visualisation de la loi normale

1. Processus stochastiques
- 2. Nombreuses données**
3. Dires d'expert
4. Loïs d'incertitude

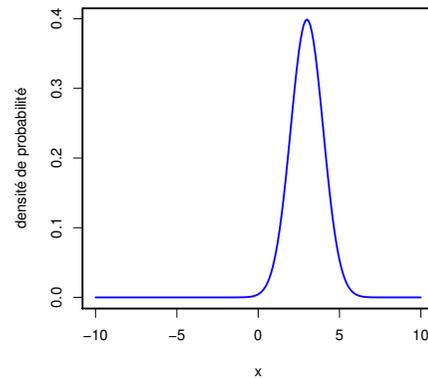
**N(0,1)**



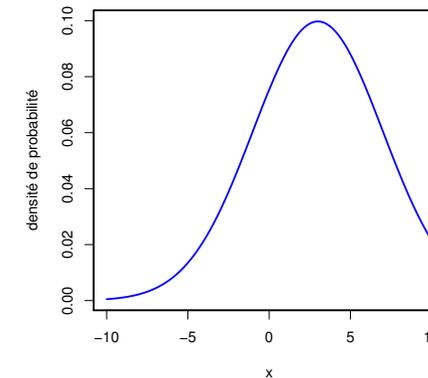
**N(0,4)**

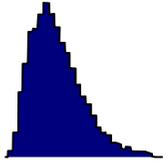


**N(3,1)**



**N(3,4)**

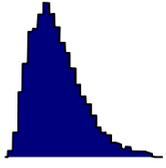




# Utilisation de la loi normale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Loi symétrique très couramment utilisée partiellement en raison du théorème de l'approximation normale
- Loi définie sur  $[-\infty; +\infty]$ : **ATTENTION** à son usage inapproprié pour des variables par nature positives (poids, taille, concentration...) lorsque le coefficient de variation dépasse 0.3.  
(coefficient de variation = écart type / moyenne)



## Description de la loi log-normale

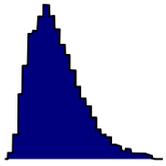
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

$$X \sim LN(\mu, \sigma)$$



$$\ln[X] \sim N(\mu, \sigma)$$

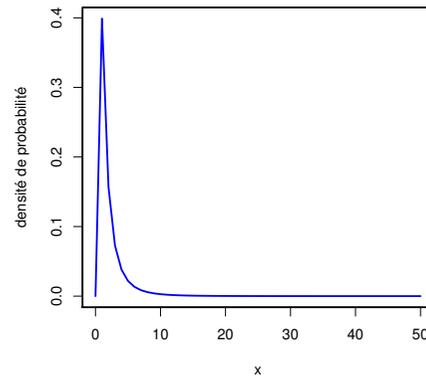
```
rlnorm(n=nb_ iterations, meanlog, sdlog)
```



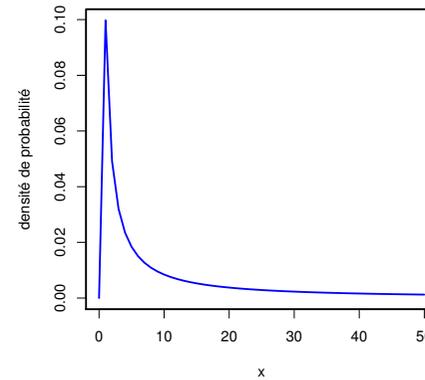
# Visualisation de la loi log-normale

1. Processus stochastiques
- 2. Nombreuses données**
3. Dires d'expert
4. Loïs d'incertitude

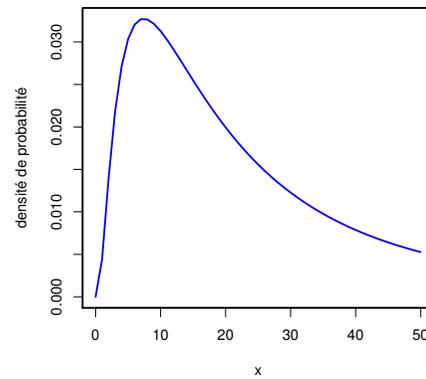
**LN(0,1)**



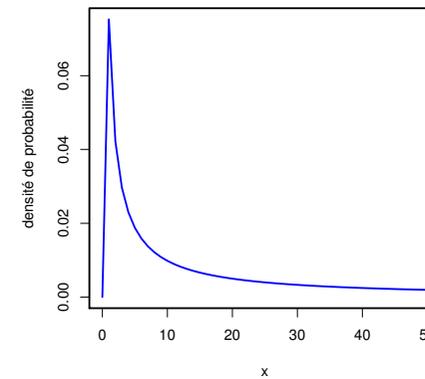
**LN(0,4)**

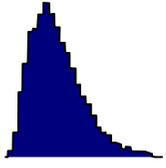


**LN(3,1)**



**LN(3,4)**

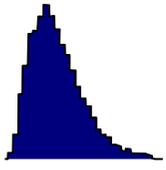




## Utilisation de la loi log-normale

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Loi très dissymétrique à droite (queue de distribution à droite pouvant être très importante) très souvent utilisée
- Adaptée aux variables auxquelles on pense en terme d'ordre de grandeur (ex.  $10^2$ ,  $10^3$ ,  $10^6$  ...),  
comme une concentration bactérienne.



# Description de la loi Bêta

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

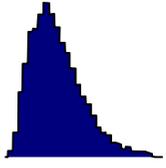
■ Ecriture :  $X \sim \text{Bêta}(\alpha_1, \alpha_2)$

■ Espérance :  $\frac{\alpha_1}{\alpha_1 + \alpha_2}$

```
rbeta(n=nb_iterations,  
      shape1= $\alpha_1$ , shape2= $\alpha_2$ )
```

■ Variance :  $\frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}$

■ Densité :  $f(x) = \frac{x^{\alpha_1 - 1} (1 - x)^{\alpha_2 - 1}}{\int_0^1 t^{\alpha_1 - 1} (1 - t)^{\alpha_2 - 1} dt}$



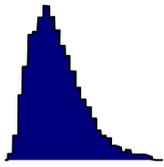
## Estimation des paramètres de la loi Bêta

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

par la méthode des moments :

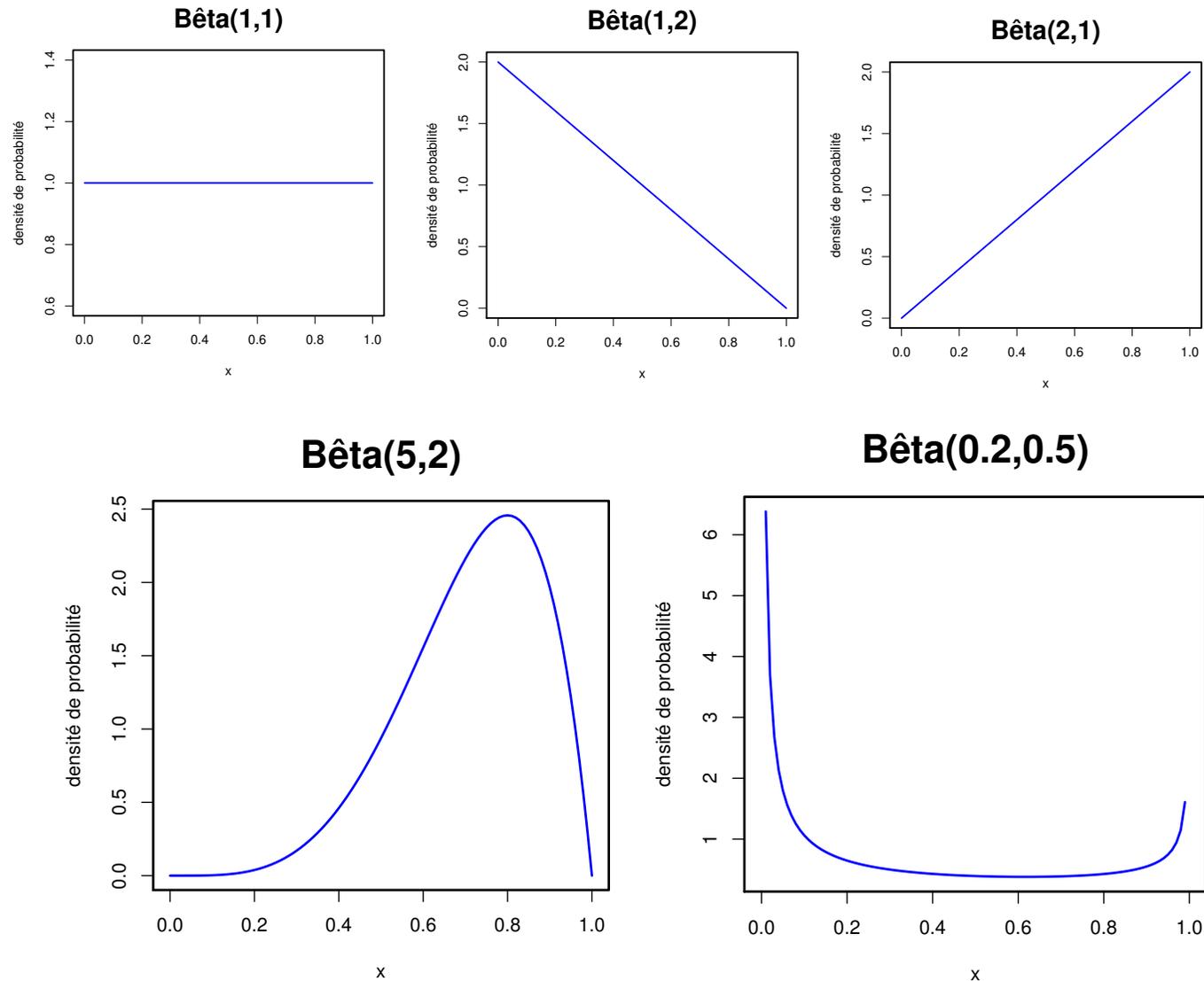
$$\hat{\alpha}_1 = \frac{\bar{x}(\bar{x} - \bar{x}^2 - \text{Var}(x))}{\text{Var}(x)}$$

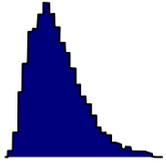
$$\alpha_2 = \hat{\alpha}_1 \left( \frac{1 - \bar{x}}{\bar{x}} \right)$$



# Visualisation de la loi Bêta

1. Processus stochastiques
- 2. Nombreuses données**
3. Dires d'expert
4. Lois d'incertitude





# Utilisation de la loi Bêta

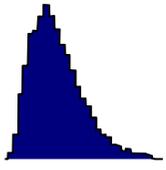
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Loi flexible définie sur  $[0;1]$
- Loi symétrique si  $\alpha_1 = \alpha_2$
- Loi permettant aussi de modéliser une variable définie sur un intervalle  $[a;b]$  en posant :

$$x = a + \text{Bêta}(\alpha_1, \alpha_2)(b - a)$$

**Explorez avec R les différentes formes que l'on obtient avec cette loi en faisant varier  $\alpha_1$  et  $\alpha_2$**

```
x <- seq(0, 1, 0.01)
plot(x, dbeta(x, shape1 = alpha1, shape2 = alpha2), type="l")
```

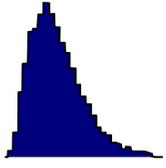


## Description de la loi de Weibull

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Ecriture :  $X \sim \text{Weibull}(\alpha, \beta)$   
 $\alpha$  paramètre de forme,  $\beta$  paramètre d'échelle
- Espérance et variance : sans expression simple
- Densité :  $f(x) = \alpha \beta^{-\alpha} x^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^\alpha\right)$
- Propriétés :  $\text{Weibull}(1, \beta) = \text{Expo}(\beta)$   
loi très proche de la loi normale lorsque  $\alpha$  est proche de 3

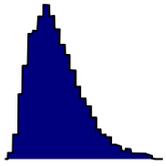
```
rweibull(n=nb_iterations, shape=alpha, scale=beta)
```



## Utilisation de la loi de Weibull

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

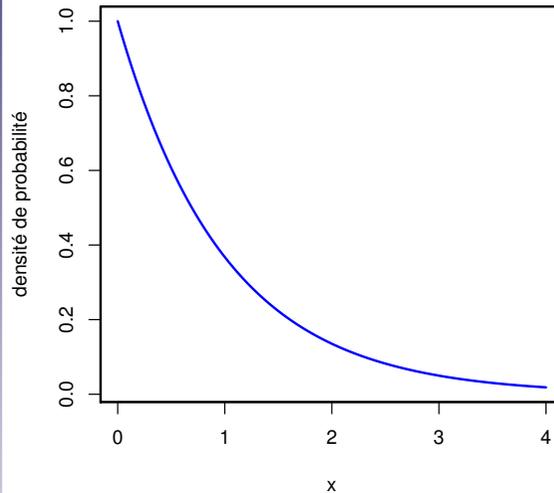
- Souvent utilisée pour décrire le temps nécessaire à l'observation d'un évènement donné lorsque la probabilité de cet évènement varie dans le temps (ex. : temps de survie avec taux de mortalité instantanée  $z(t)$  diminuant ou augmentant en fonction du temps :  $z(t) = \lambda^\alpha \times \alpha \times t^{\alpha-1}$  ).
- **Loi flexible**, proche de la loi Gamma en un peu moins dissymétrique (queue à droite moins importante)  
et autorisant même une dissymétrie à gauche lorsque  $\alpha$  est grand ( $> 3.6$ )



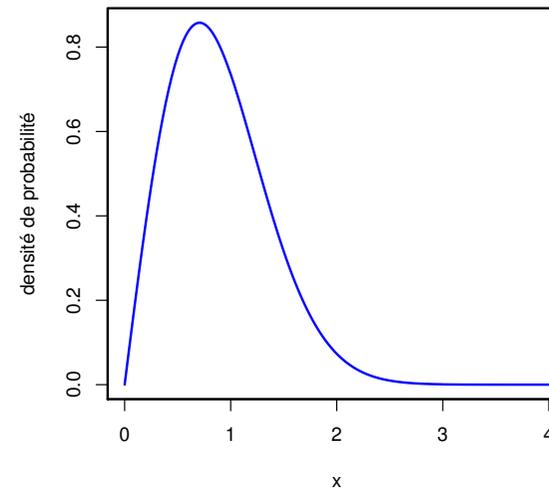
# Visualisation de la loi de Weibull

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

Weibull(1,1)

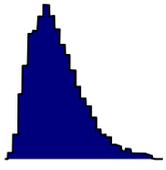


Weibull(2,1)



Explorez avec R les différentes formes que l'on obtient avec cette loi en faisant varier son paramètre de forme  $\alpha$

```
x <- seq(0,10,0.1)
plot(x,dweibull(x,shape = alpha,scale = 5),type="l")
```



# Description de la loi logistique

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Ecriture :  $X \sim \text{logistique}(\alpha, \beta)$   
 $\alpha$  paramètre de localisation,  $\beta$  paramètre d'échelle

- Espérance :  $\alpha$

- Variance : 
$$\frac{\beta^2 \times \pi^2}{3}$$

- Fonction de répartition  
logistique :

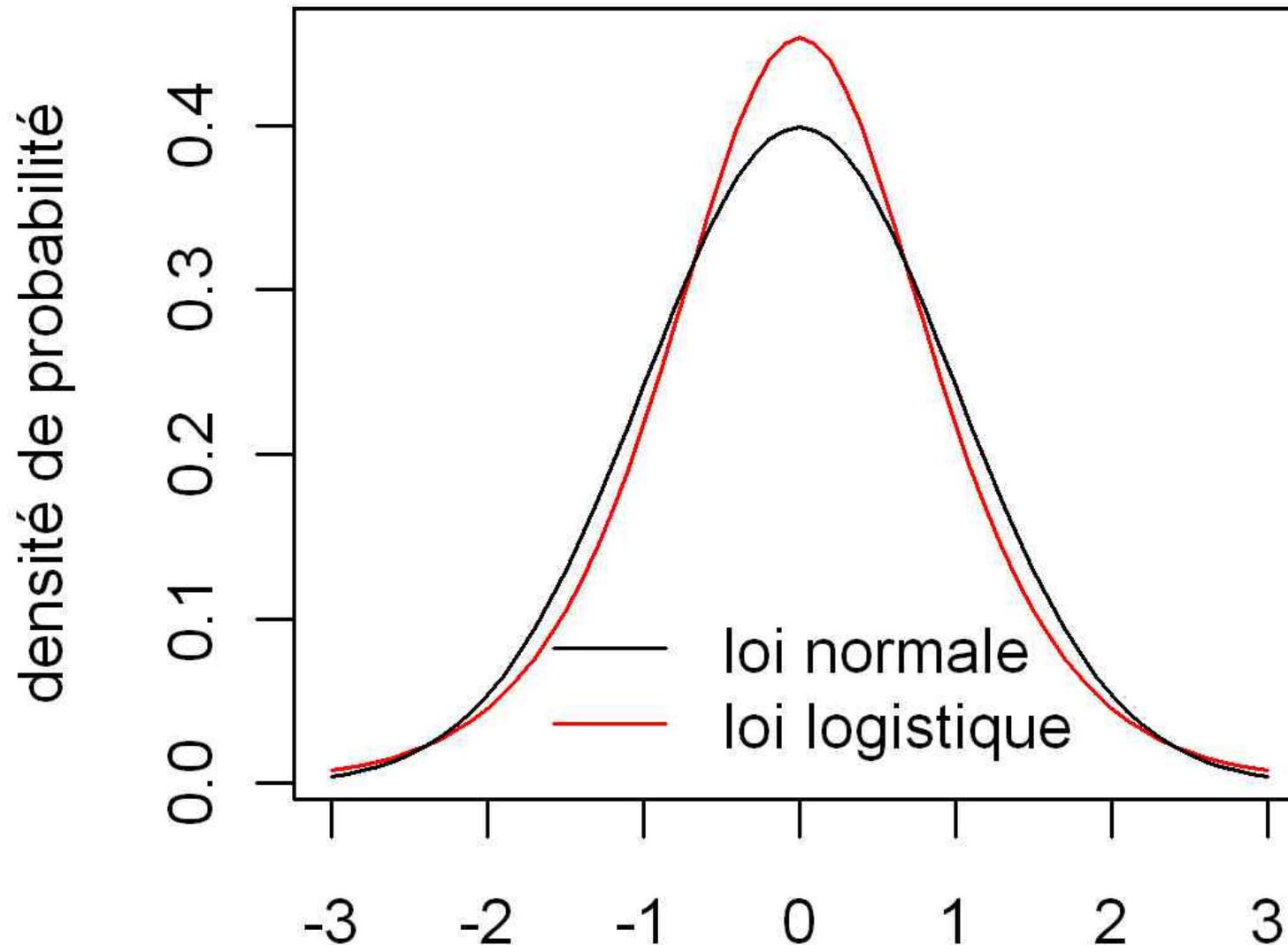
$$F(x) = \frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}$$

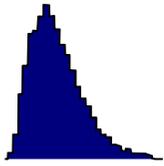
- Propriétés :

Loi symétrique avec un coefficient de pointicité (kurtosis) plus élevée que la loi normale (loi **leptokurtique** : plus pointue avec queues plus épaisses)

```
rlogis(n = nb_tirages, location =  $\alpha$ , scale =  $\beta$ )
```

## Loi logistique comparée à la loi normale de même moyenne et écart type

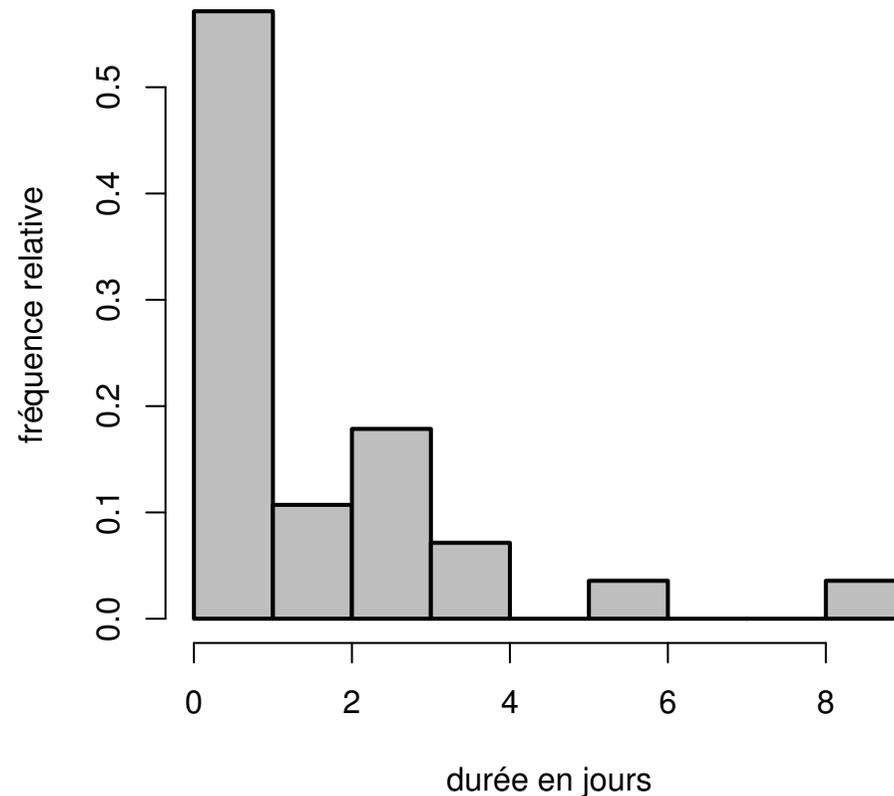




## Ex. données de durée de stockage au frais

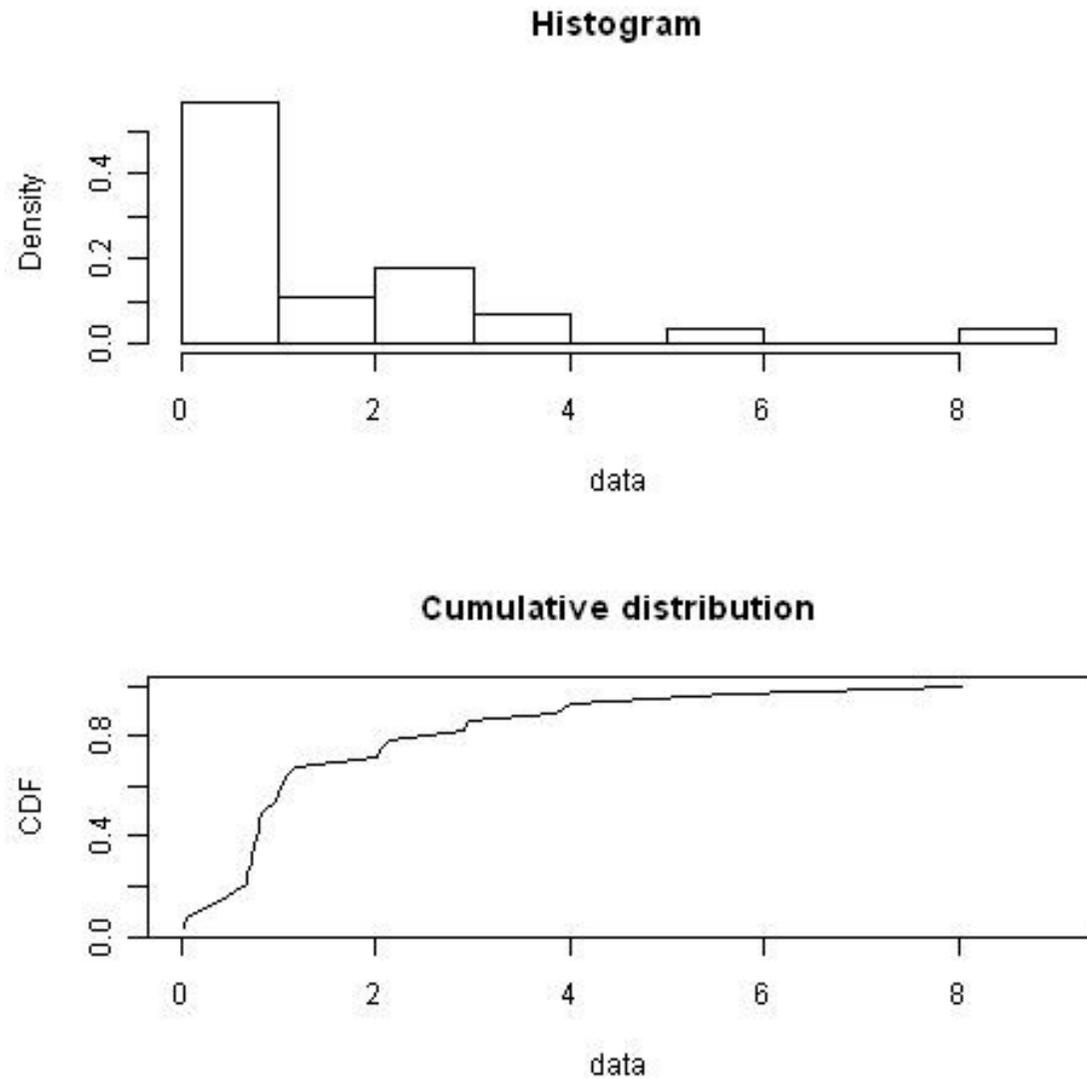
1. Processus stochastiques
2. **Nombreuses données**
3. Dires d'expert
4. Loïs d'incertitude

- Durée de stockage dans le frigo d'un produit alimentaire donné avant consommation (observations lors d'une enquête de consommation)



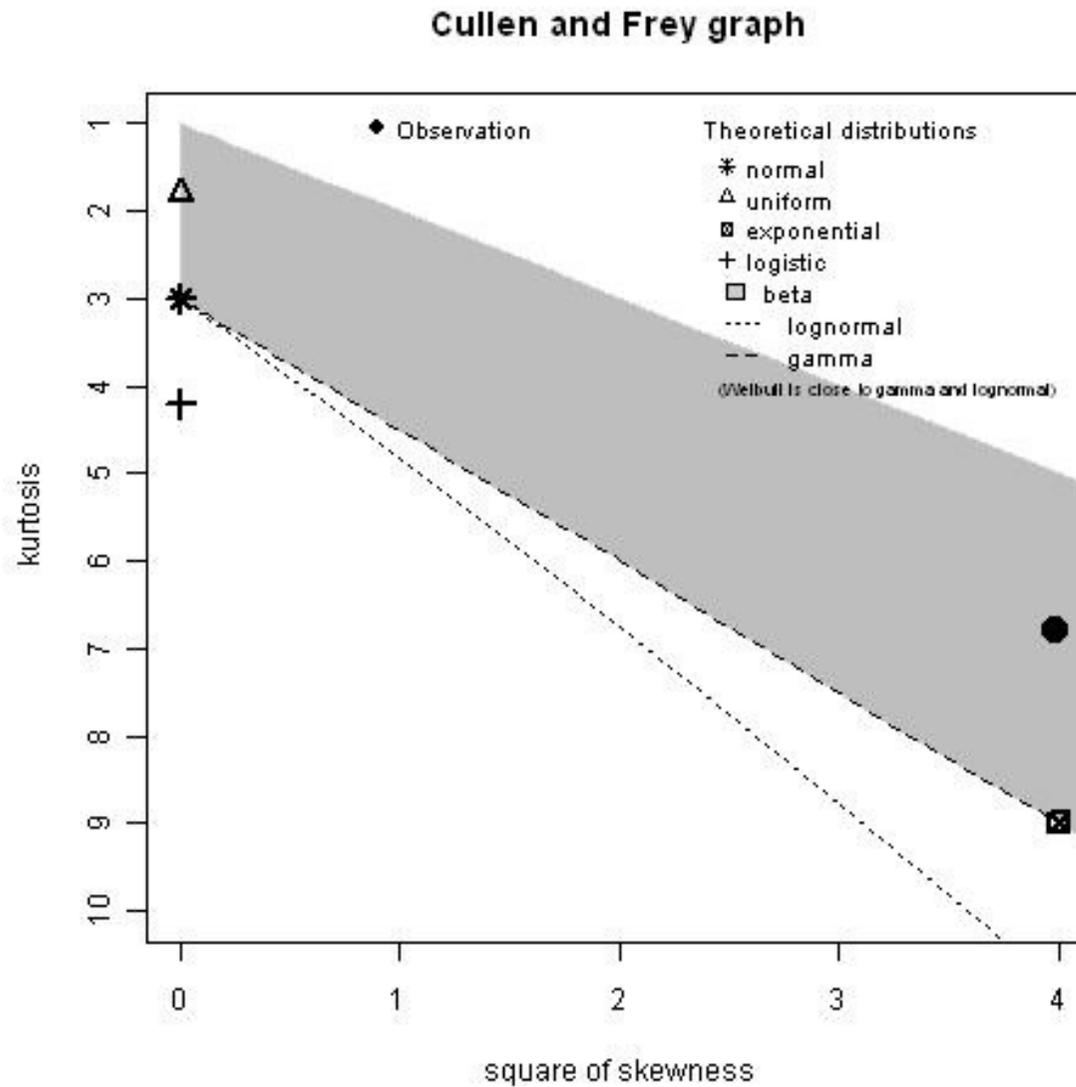
# Utilisation de la fonction `plotdist (fitdistrplus)`

```
plotdist (data=vecteur_de_données)
```



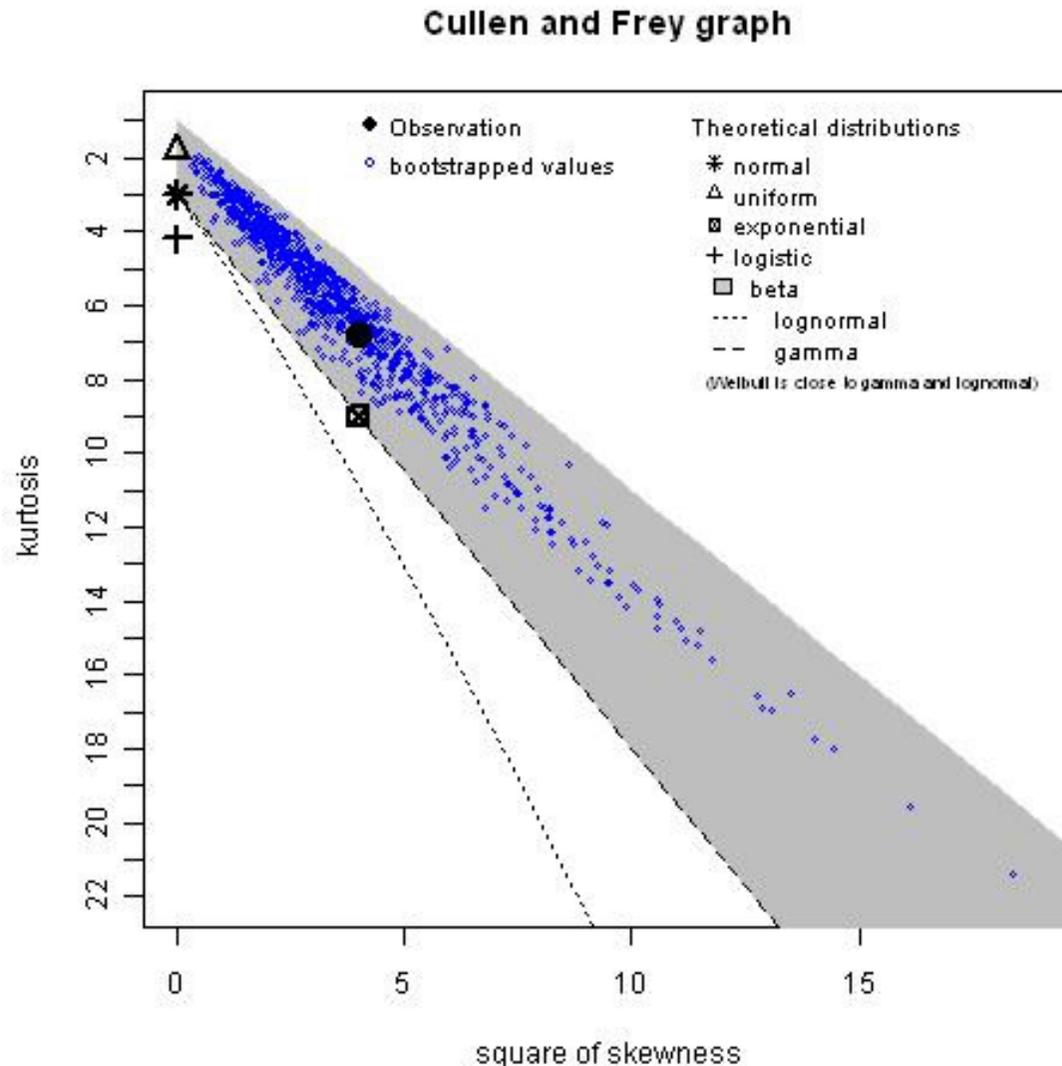
# Utilisation de la fonction `descdist` (`fitdistrplus`)

```
descdist (data=vecteur_de_données)
```



# Utilisation de la fonction `descdist` (`fitdistrplus`)

```
descdist (data=vecteur_de_données, boot=1000)
```

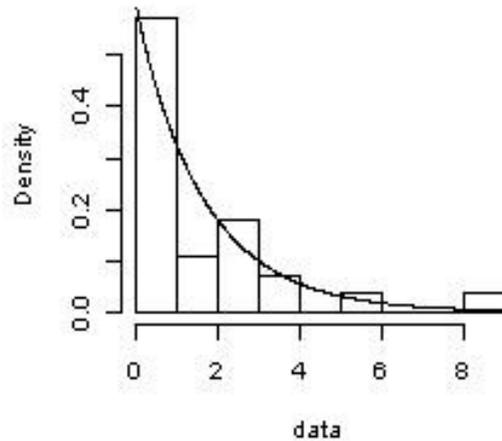


**Ajout de  
l'incertitude  
sur les 2  
coefficients**

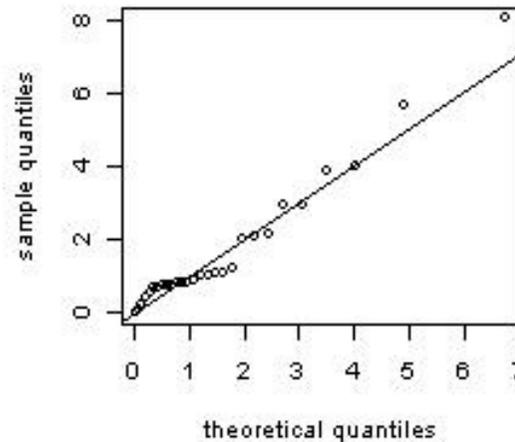
# Utilisation de la fonction `fitdist` (`fitdistrplus`)

```
f <- fitdist(data = vecteur_de_données,  
             distr = distribution)
```

Empirical and theoretical distr.

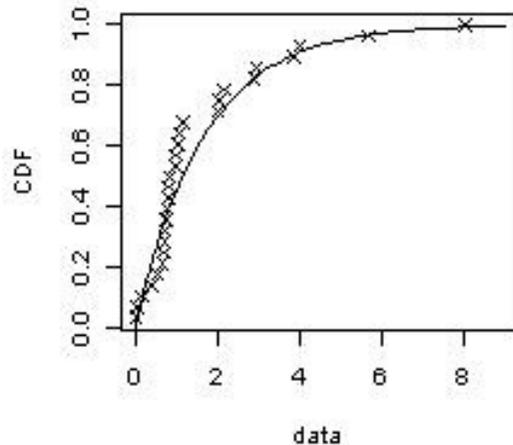


QQ-plot

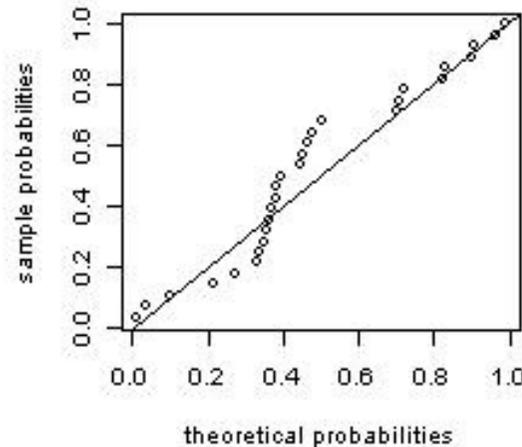


**Ex.:**  
**ajustement**  
**d'une loi**  
**gamma**

Empirical and theoretical CDFs



PP-plot



`plot(f)`

## Utilisation de la fonction `fitdist` (`fitdistrplus`)

### Ex.: ajustement d'une loi gamma

```
Fitting of the distribution ' gamma ' by
maximum likelihood
Parameters :
      estimate Std. Error
shape    0.991    0.233
rate     0.592    0.179
Loglikelihood:  -42.4    AIC:  88.8    BIC:  91.5
Correlation matrix:
      shape  rate
shape 1.000  0.778
rate  0.778  1.000
```

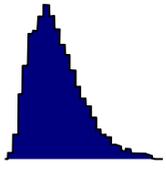
**summary (f)**

Utilisation de la fonction `fitdist` (`fitdistrplus`)

**Ex.: ajustement d'une loi gamma**  
**Statistiques d'ajustement**

```
Kolmogorov-Smirnov statistic: 0.172  
Cramer-von Mises statistic: 0.157  
Anderson-Darling statistic: 0.771
```

**gofstat (f)**



## Ex. résultats de quelques ajustements

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

### Statistiques de Anderson-Darling

**Plus elle est faible, meilleur est l'ajustement.**

- Loi log-normale : 1.13
- Loi Gamma : 0.771
- **Loi de Weibull : 0.767**
- Loi exponentielle : 0.773

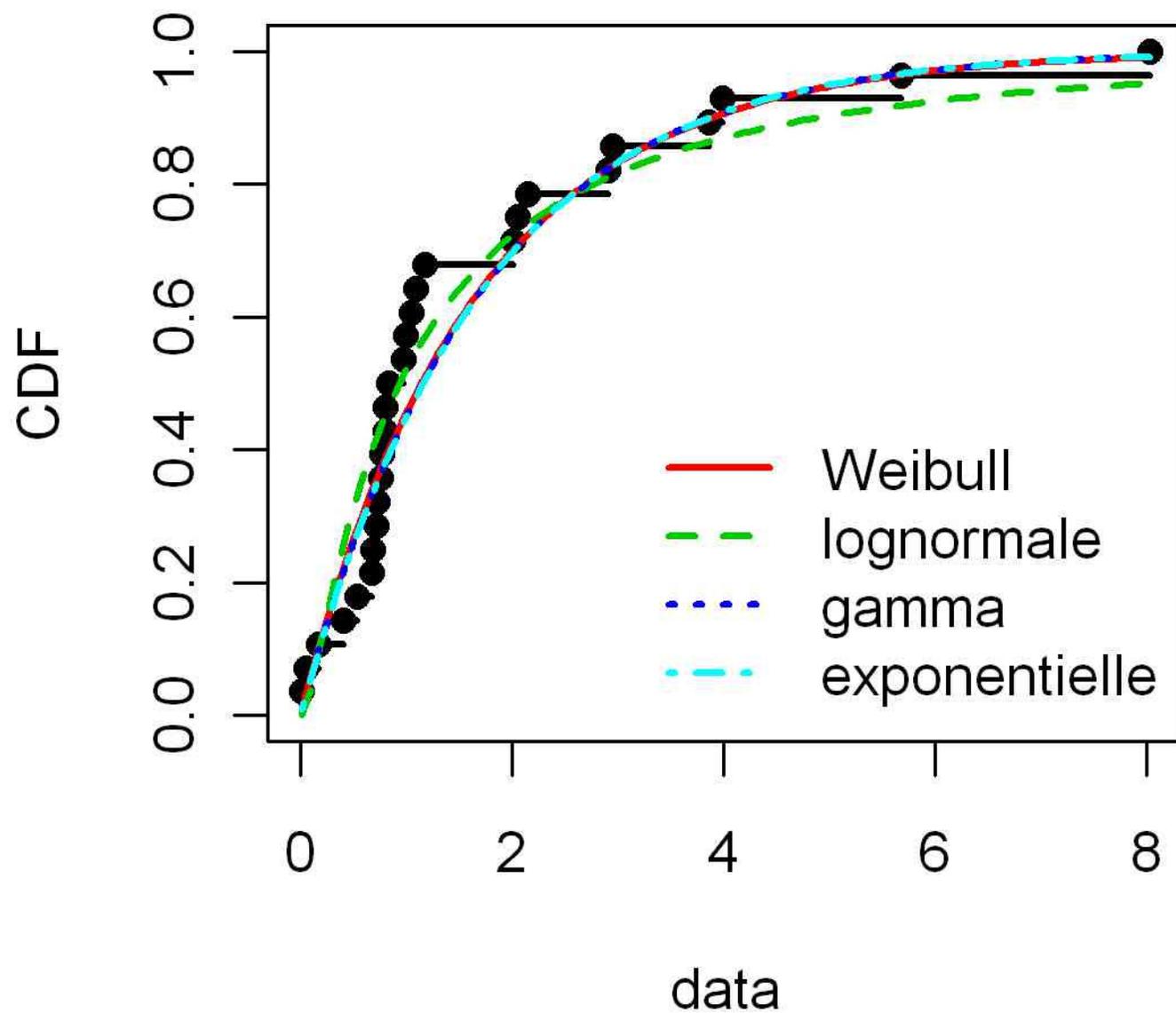
meilleur ajustement : Weibull

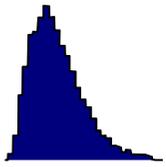
**mais 3 dernières lois très comparables**

→ **choix de la loi exponentielle** qui comporte un seul paramètre (**parcimonie**)

- AIC : gamma 89, Weibull 89, **expo 87**
- BIC : gamma 91, Weibull 91, **expo 88**

## Comparaison graphique des lois ajustées

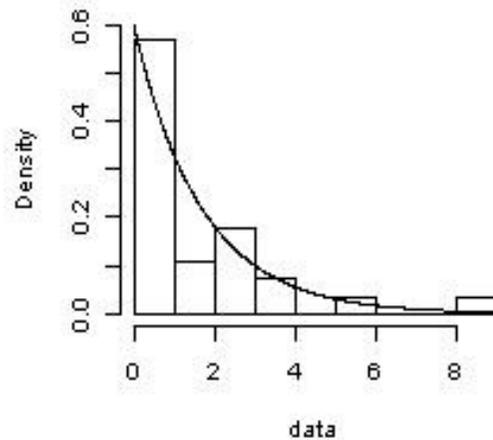




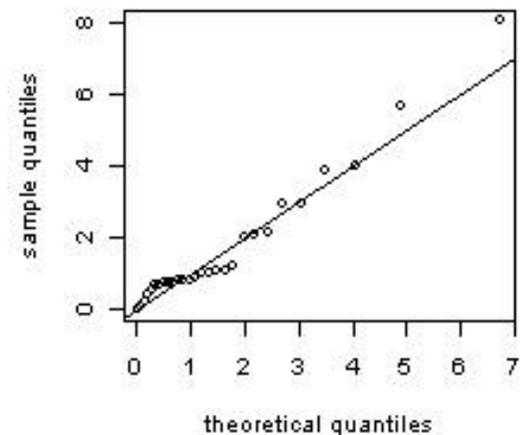
# Ex. loi expo ajustée aux données

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

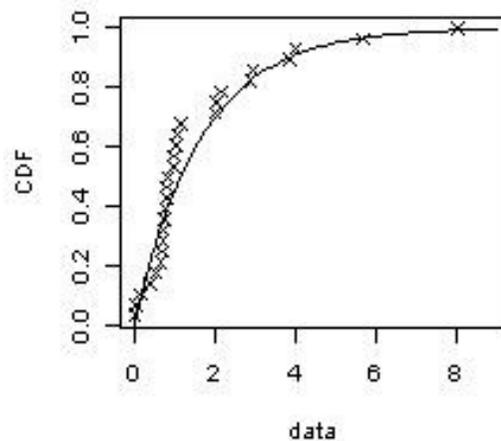
Empirical and theoretical distr.



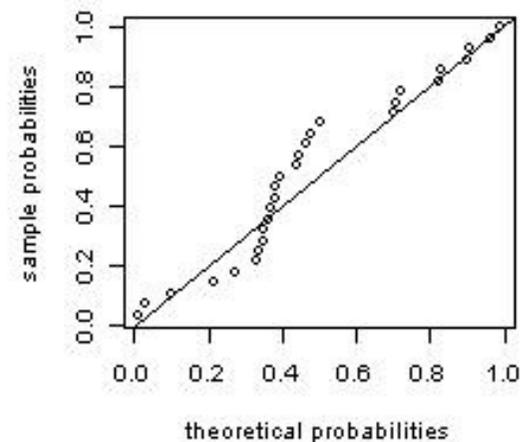
QQ-plot



Empirical and theoretical CDFs



PP-plot



Pour vous entraîner, réalisez le même type de travail sur le jeu de données `groundbeef` du package `fitdistrplus`

```
data(groundbeef)
str(groundbeef)
s <- groundbeef$serving
plotdist(s)
descdist(s, boot=1000)
fG <- fitdist(s, "gamma")
plot(fG); summary(fG); gofstat(fG)
fW <- fitdist(s, "weibull")
plot(fW) ; summary(fW); gofstat(fW)
fLN <- fitdist(s, "lnorm")
plot(fLN) ; summary(fLN); gofstat(fLN)
cdfcomp(list(fW, fLN, fG), xlab="serving sizes (g)",
            legendtext=c("Weibull", "lognormal", "gamma"))
```

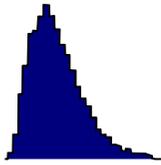
## Puis simulez des échantillons à partir des trois méthodes envisagées

```
# pour faire 3 figures sur la même fenêtre
par(mfrow=c(1,3))

# simulation par tirage avec remise dans les données
hist(sample(s,size=1000,replace=TRUE))

# simulation à partir de la fonction de répartition
# interpolée après ajout d'une information d'expert
# sur les valeurs min et max afin de compléter
# la définition de cette fonction
hist(rempiricalC(n=1000, min=0, max=250,values=s))

# simulation par tirage
# dans la loi gamma ajustée aux données
hist(rgamma(n=1000,shape=fG$estim["shape"],
            rate=fG$estim["rate"]))
```



# Cas des données semi-quantitatives

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

## Censures fréquentes dans les données

Exemple tiré de *Busschaert et al. 2010*, (*Listeria monocytogenes* dans du saumon fumé en distribution en Belgique)

- **Données détection :**  
 $< 0.04$  : absence dans 25 g  
 $< 1$  : absence dans 1 g
- **Données dénombrement :**  
 $< 100$  : absence dans 0.01 g  
 Étalement de 0.1 ml du bouillon dilué au 1/10  
 $< 10$  : absence dans 0.1 g

Nombre d'échantillons	Concentration (UFC.g <sup>-1</sup> )
54	$< 0.04$
2	$< 100$
26	0.04 - 10
1	15
8	0.04 - 100
2	$> 100$
1	$< 1$
1	$> 1$
7	0.04 - 1
1	1 - 100

## Ajustement d'une loi par maximum de vraisemblance

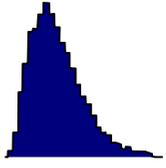
## Maximum de vraisemblance pour données censurées

Pour une distribution de paramètre  $\theta$  on choisit  $\theta$  qui maximise la vraisemblance des données

- $x$  (observation non censurée)
- $x_u$  (observation censurée à gauche :  $< x_u$ )
- $x_l$  (observation censurée à droite :  $> x_l$ )
- $x_l ; x_u$  (définie par un intervalle :  $> x_l$  et  $< x_u$ )

$$\max \left( \prod_{i=1}^n \Pr(x_i | \theta) \right) =$$

$$\max \left( \prod f(x) \times \prod F(x_u) \times \prod (1 - F(x_l)) \times \prod (F(x_u) - F(x_l)) \right)$$



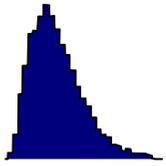
## Codage des données avant analyse

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

Sous forme d'un objet R de type  
`data.frame` à 2 colonnes

- `left` : limite inférieure ou NA si censure à gauche ( $x < \text{valeur}$ )
- `right` : limite supérieure ou NA si censure à droite ( $x > \text{valeur}$ )

`left = right` pour les observations non censurées

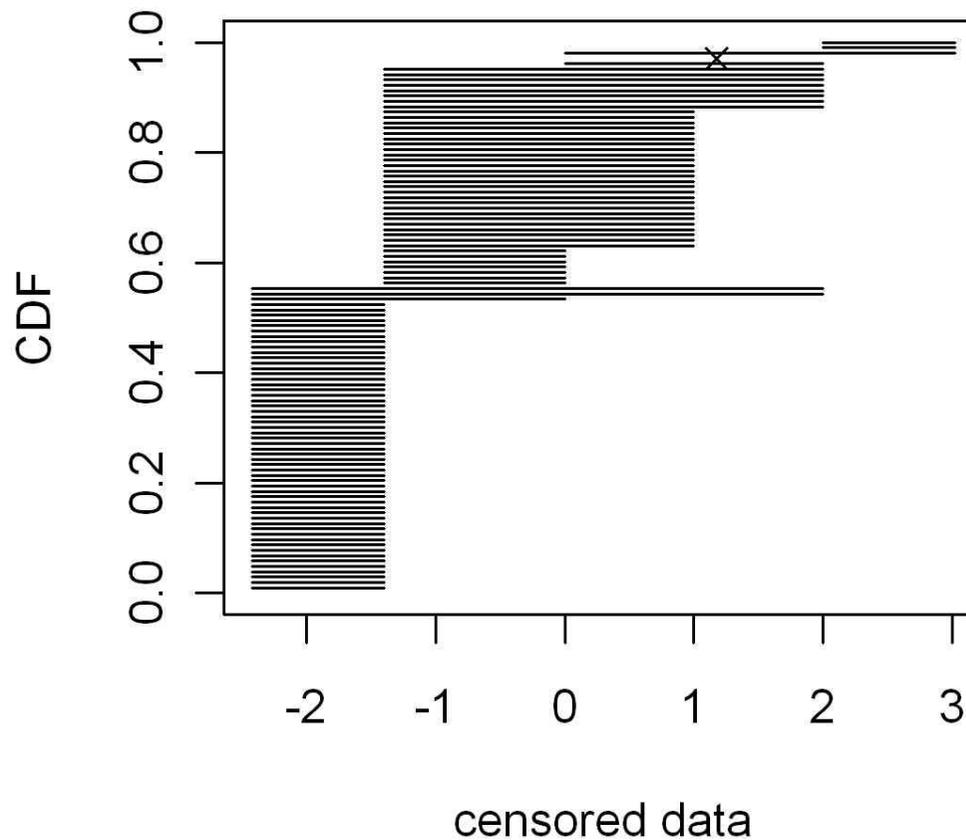


# Représentation des données (ici en $\log_{10}$ )

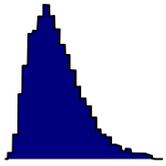
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

```
plotdistcens(data = jeu_de_données, Turnbull=FALSE)
```

## Cumulative distribution



Représentation  
des données  
censurées par  
des intervalles

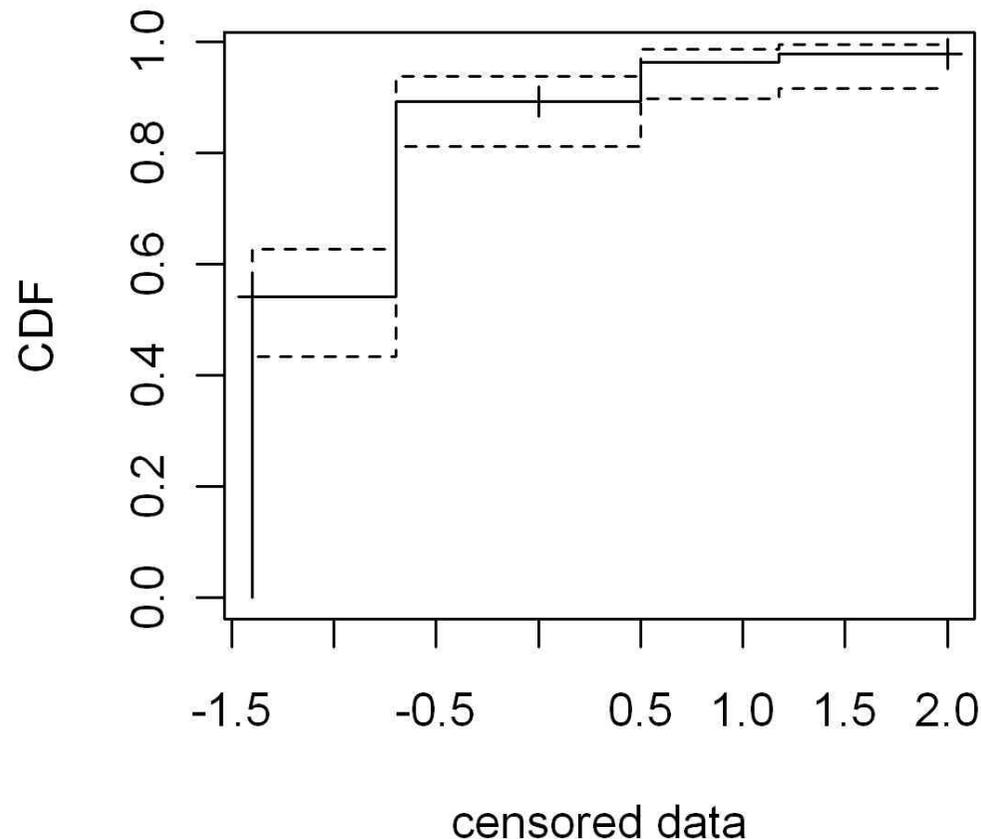


## Autre représentation des données (ici en $\log_{10}$ )

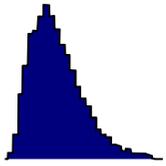
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

```
plotdistcens(data = jeu_de_données)
```

### Cumulative distribution



Représentation des  
données censurées  
par estimation non  
paramétrique de la  
courbe de fréquences  
cumulées et de sa  
bande de confiance à  
95%  
(méthode de Turnbull)

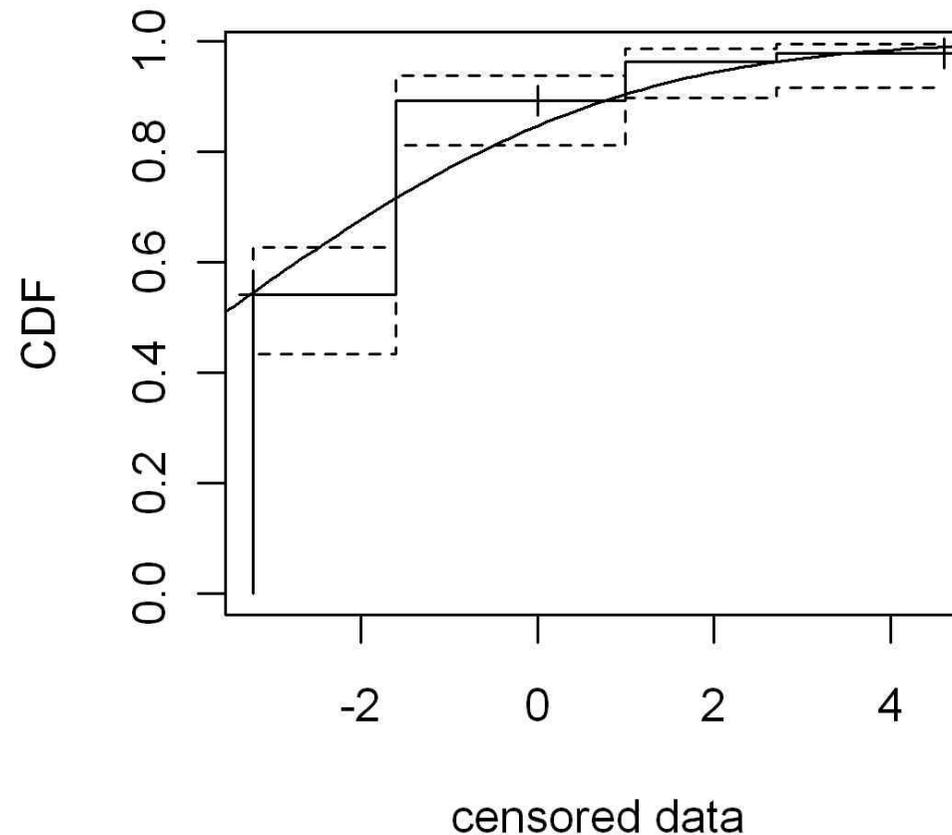


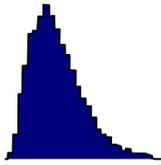
# Ajustement d'une loi normale sur les données en $\log_{10}$

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

```
fitdistcens(data = jeu_de_données, distr = distribution)
```

Cumulative distribution



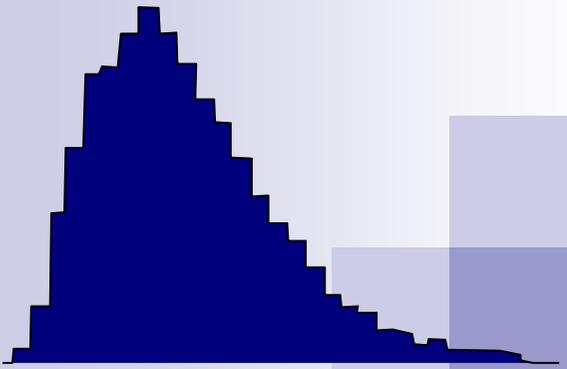


## A retenir



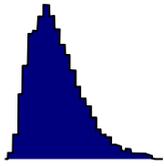
1. Processus stochastiques
2. **Nombreuses données**
3. Dires d'expert
4. Lois d'incertitude

- Description non paramétrique
  - Tirage avec remise dans les données
  - Tirage à partir de la fonction de répartition interpolée
- Description paramétrique
  - Sélection de lois candidates
  - Ajustement d'une ou plusieurs lois
  - Comparaison des ajustements
- Cas particulier des données censurées : ajustement d'une loi paramétrique par maximum de vraisemblance recommandé sur données correctement codées



### 3. On doit se baser sur des dires d'experts

Données absentes ou trop peu  
nombreuses pour caractériser une  
loi



## Deux situations

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

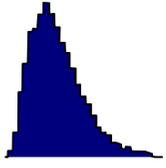
### ■ Loi de forme connue :

- Interrogation des experts pour fixer les paramètres de cette distribution souvent à partir de quelques quantiles

### ■ Loi de forme inconnue :

- Utilisation d'une loi parmi les classiques (uniforme, triangulaire, Pert ou générale)
- Interrogation des experts pour fixer les paramètres (min, max, mode, quantiles...)

L'interrogation des experts est délicate :  
des méthodes sophistiquées existent



# Loi uniforme

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

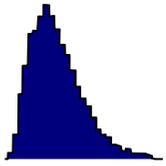
■ Ecriture :  $X \sim \text{Unif}(min, max)$

■ Espérance :  $\frac{min + max}{2}$

■ Variance :  $\frac{(max - min)^2}{12}$

■ Densité :  $f(x) = \frac{1}{max - min}$

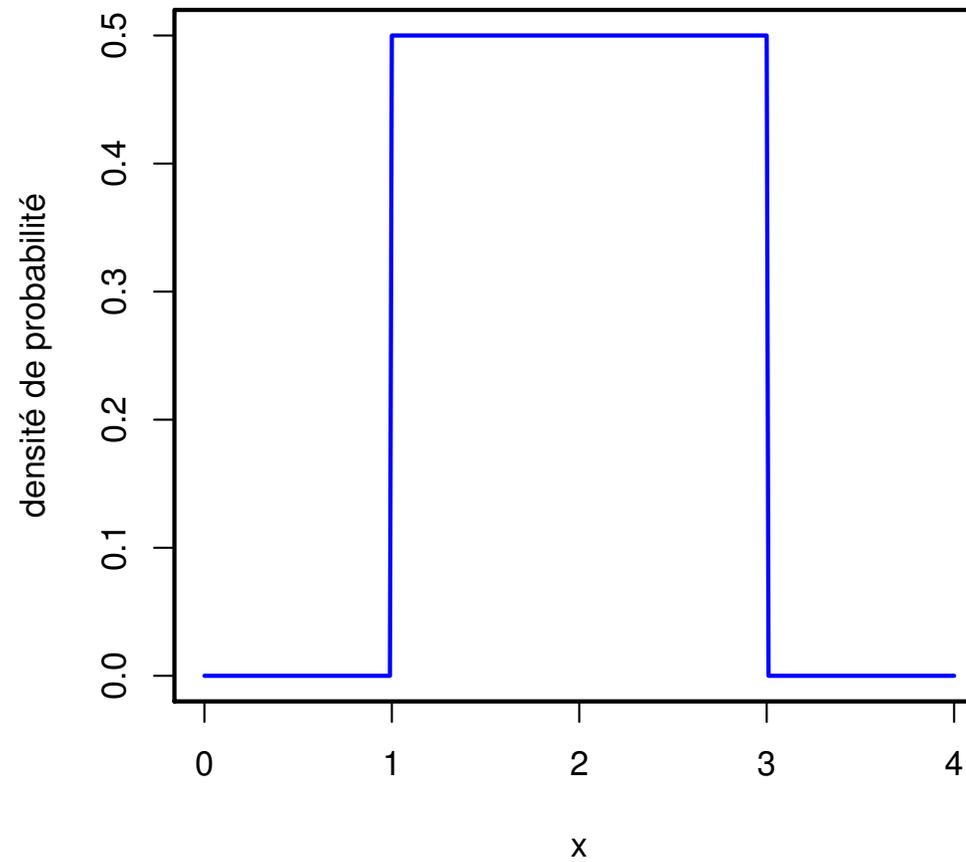
```
runif(n=nb_iterations, min, max)
```

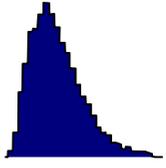


# Visualisation de la loi uniforme

1. Processus stochastiques
2. Nombreuses données
3. **Dires d'expert**
4. Loïs d'incertitude

## Unif(1,3)

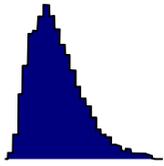




## Utilisation de la loi uniforme

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Les experts proposent des valeurs minimale et maximale
- Les experts affirment que toutes les valeurs situées dans l'intervalle sont équiprobables
- Cette distribution n'est pas « naturelle » et très sensible aux valeurs min et max



# Loi triangulaire

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

■ Ecriture :  $X \sim \text{Triang}(\min, \text{mode}, \max)$

■ Espérance :  $\frac{\min + \text{mode} + \max}{3}$

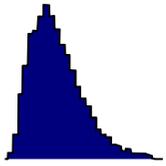
```
rtriang(n=nb_iterations, min, mode, max)
```

■ Variance :  $\frac{\min^2 + \text{mode}^2 + \max^2 - \min \times \text{mode} - \min \times \max - \max \times \text{mode}}{18}$

■ Densité :

$$f(x) = \frac{2(x - \min)}{(\text{mode} - \min)(\max - \min)} \quad \text{si } \min \leq x \leq \text{mode}$$

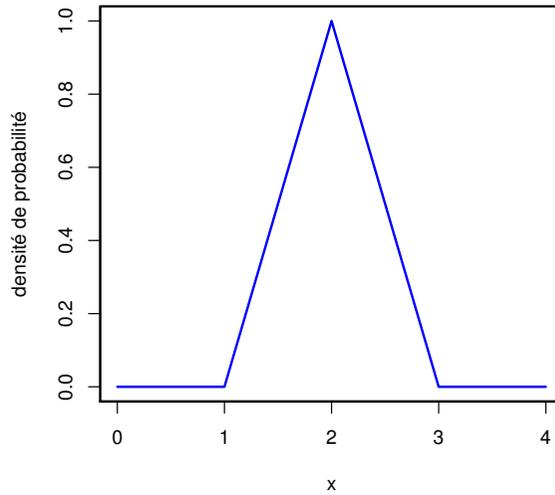
$$f(x) = \frac{2(\max - x)}{(\max - \min)(\max - \text{mode})} \quad \text{si } \text{mode} \leq x \leq \max$$



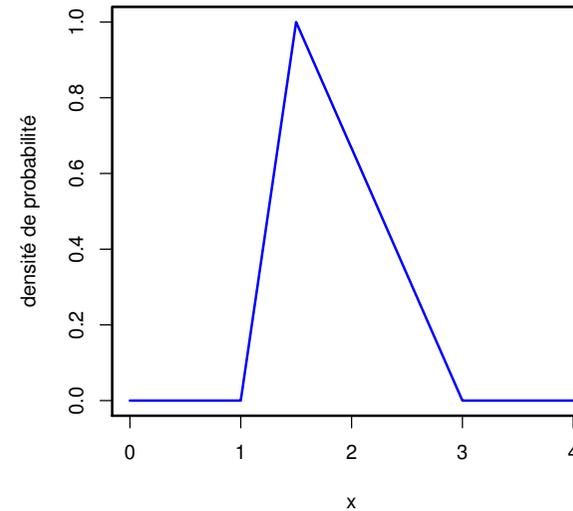
# Visualisation de la loi triangulaire

1. Processus stochastiques
2. Nombreuses données
- 3. Dires d'expert**
4. Lois d'incertitude

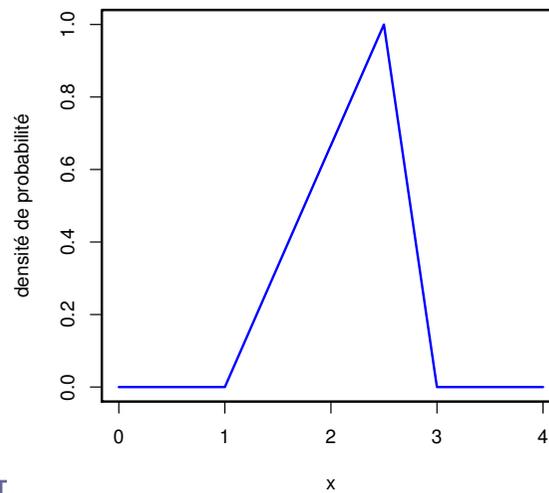
**Triang(1,2,3)**

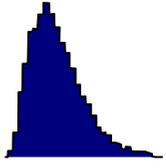


**Triang(1,1.5,3)**



**Triang(1,2.5,3)**



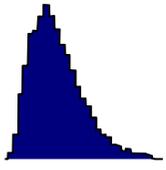


## Utilisation de la loi triangulaire

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Les experts proposent des valeurs minimale et maximale et plus probable
- Cette distribution n'est pas « naturelle » et assez sensible aux valeurs min, max et mode

**on lui préfère souvent la loi Pert**



# Loi Pert

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Ecriture :  $X \sim \text{Pert}(\text{min}, \text{mode}, \text{max})$

```
rpert (n=nb_iterations, min, mode, max)
```

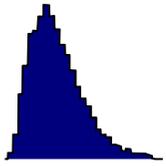
- Espérance :  $\mu = \frac{\text{min} + 4\text{mode} + \text{max}}{6}$

- Définition :  $X \sim \text{Bêta}(\alpha_1, \alpha_2) \times (\text{max} - \text{min}) + \text{min}$

avec

$$\alpha_1 = \frac{(\mu - \text{min})(2\text{mode} - \text{min} - \text{max})}{(\text{mode} - \mu)(\text{max} - \text{min})}$$

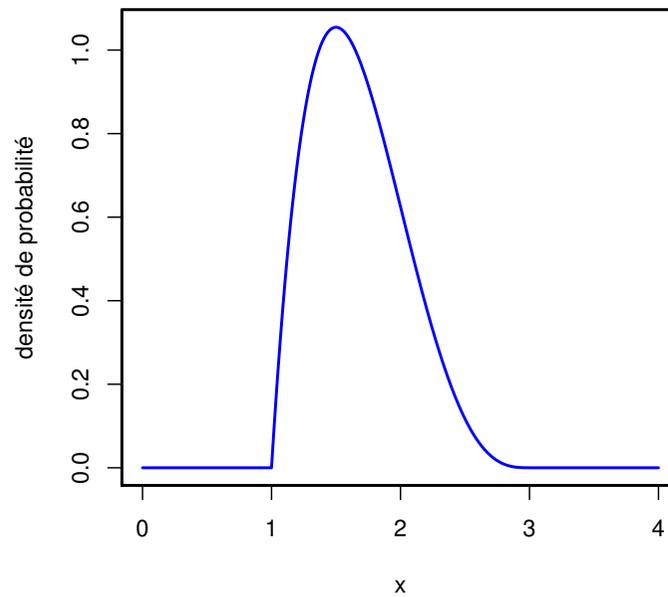
$$\alpha_2 = \frac{\alpha_1(\text{max} - \mu)}{(\mu - \text{min})}$$



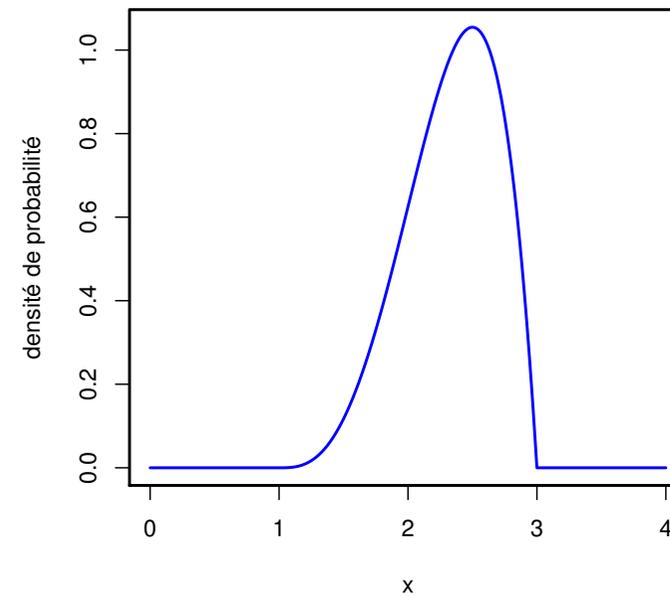
# Visualisation de la loi Pert

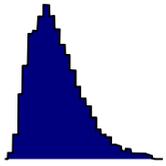
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

**Pert(1,1.5,3)**



**Pert(1,2.5,3)**

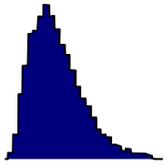




## Ex. Loi de $T_{\min}$ de *Bacillus cereus*

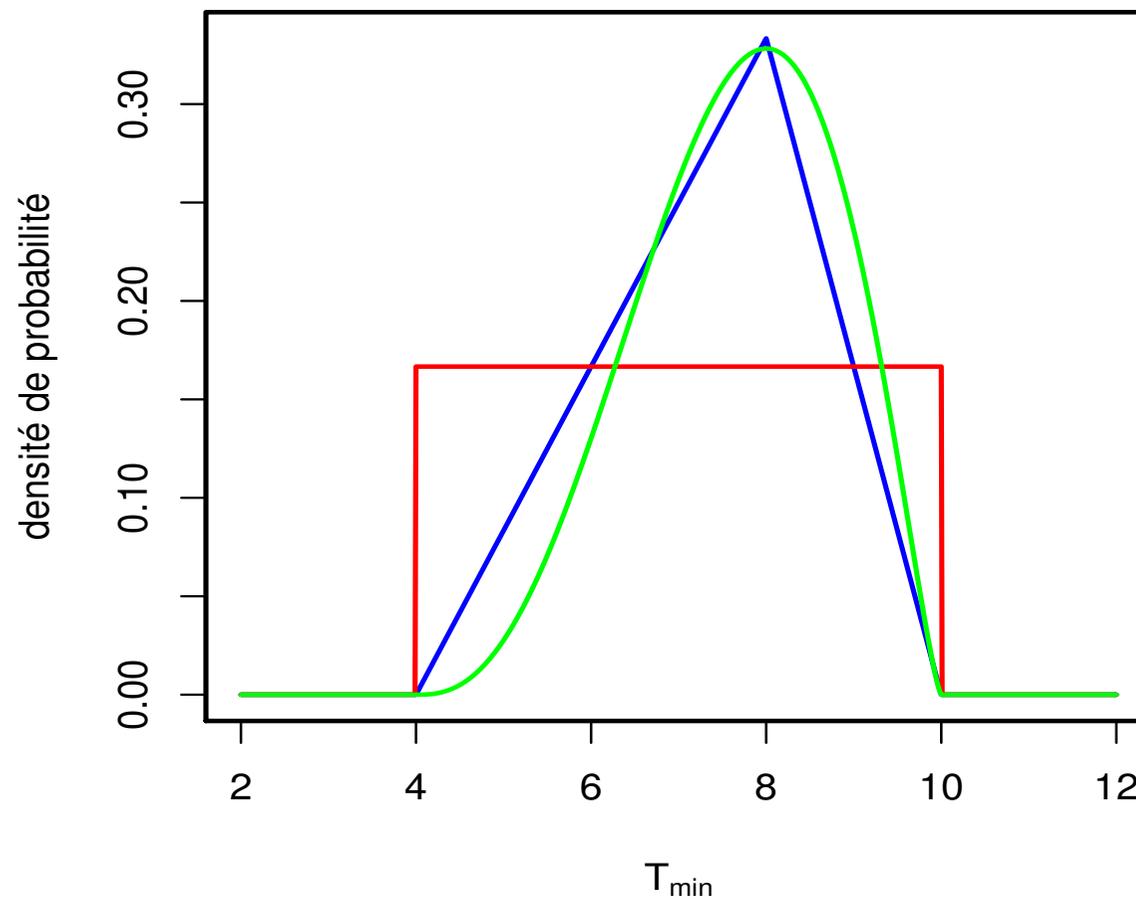
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

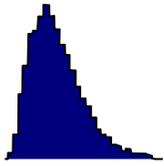
- Au sein de l'espèce *Bacillus cereus* la température minimale de croissance  $T_{\min}$  varie d'une souche à l'autre,  
d'après un groupe d'experts
  - entre 4 et 10°C
  - avec une valeur la plus probable à 8°C
- Paramétrisation des 3 lois classiques à partir de ces dires d'expert:
  - Unif(4,10)
  - Triang(4,8,10)
  - Pert(4,8,10)



## Ex. Comparaison des 3 lois décrivant la variabilité de $T_{\min}$

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude



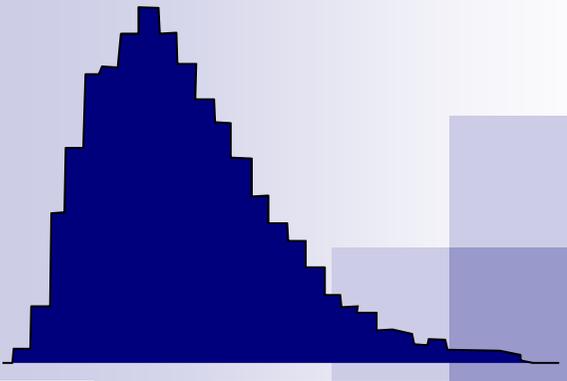


## A retenir

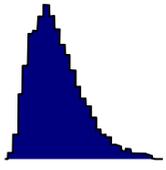


1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Loïs d'incertitude

- Paramétrisation d'une loi de forme connue à partir d'interrogations d'experts
  - Paramétrisation à partir de quelques quantiles
- Description à partir d'interrogations d'experts par une des 3 lois
  - Loi uniforme (très sensible aux min et max)
  - Loi triangulaire (sensible aux min, max et mode)
  - Loi Pert (dérivée de la loi bêta)
- Interrogation des experts = tâche délicate à effectuer avec des méthodes appropriées



## 4. Modélisation de l'incertitude



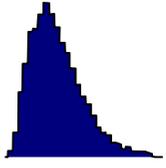
## Utilisation de la statistique fréquentiste

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Incertitude sur les paramètres classiques (moyenne, écart type, proportion) caractérisée par un **intervalle de confiance**

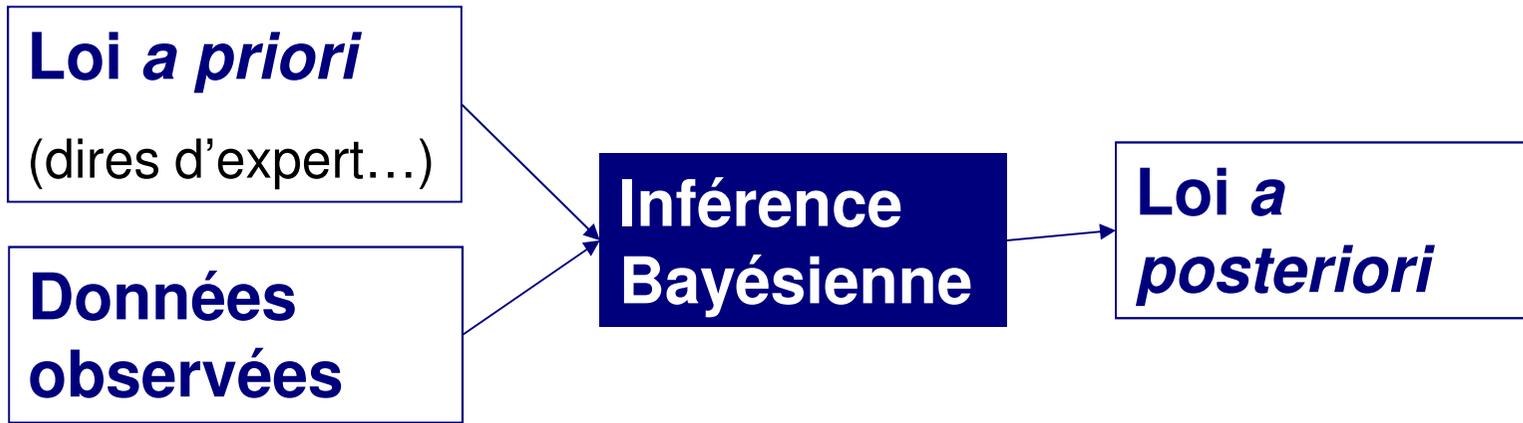
intervalle de confiance à 95% = intervalle ayant une probabilité de 0.95 de contenir la valeur vraie du paramètre

- Il n'est pas évident d'en déduire une loi d'incertitude car le paramètre est considéré comme inconnu mais fixe (non défini par une loi)



# Utilisation de la statistique Bayésienne

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude



## Utilisation de l'inférence Bayésienne

- soit avec une loi *a priori* informative
- soit avec une loi *a priori* non informative (on rejoint alors parfois la statistique fréquentiste)

# Rappel du théorème de Bayes et son application en inférence

## Théorème de Bayes

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})}$$

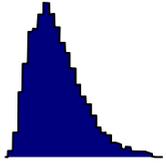
$$\Pr(\theta|X) = \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X)}$$

Loi *a posteriori*

vraisemblance

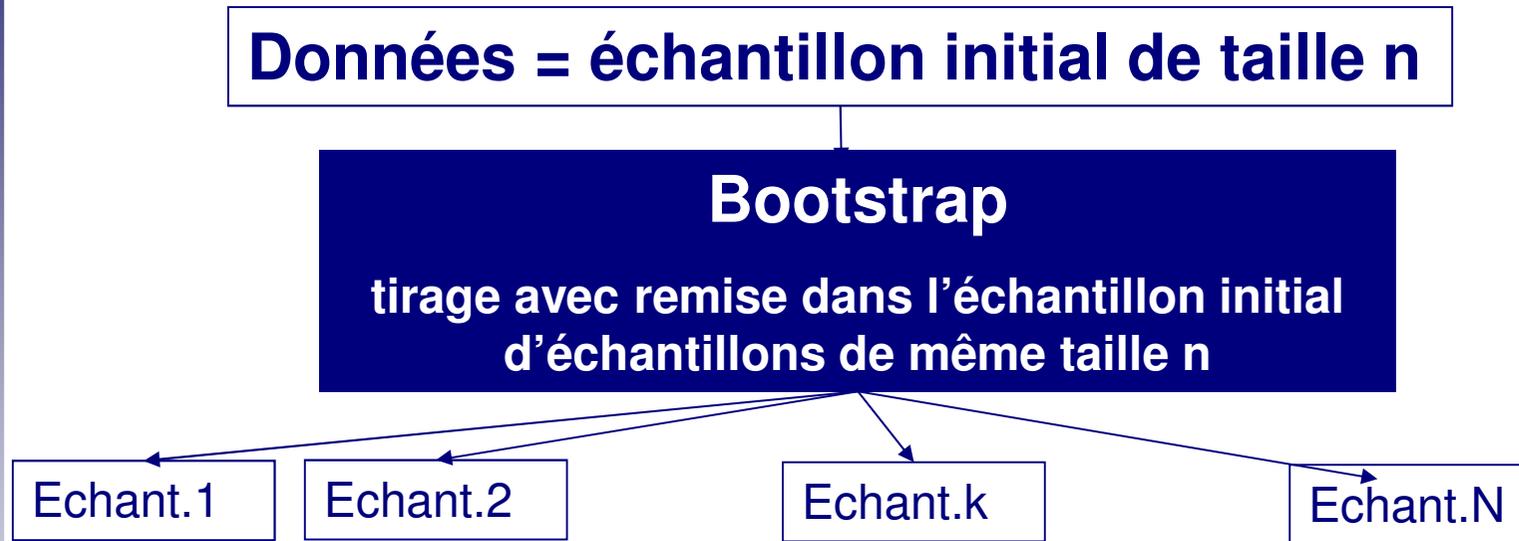
Loi *a priori*

$$\Pr(\theta|X) \propto \Pr(X|\theta)\Pr(\theta)$$

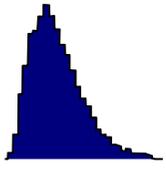


# Utilisation du bootstrap ou rééchantillonnage

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude



- $N$  valeurs estimées du paramètre  
(avec  $N$  très grand)
- loi d'incertitude empirique du paramètre
- éventuellement ajustement d'une loi paramétrée  
sur cette loi empirique

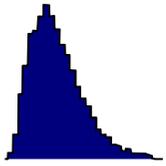


## Ex. incertitude sur une proportion

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- La proportion de lots d'un aliment donné contaminé par une bactérie pathogène donnée est estimée à 0.001 à partir d'une enquête nationale réalisée sur 4000 lots ( $n=4000$ ).
- Quelle loi utiliser pour modéliser l'incertitude sur cette prévalence estimée

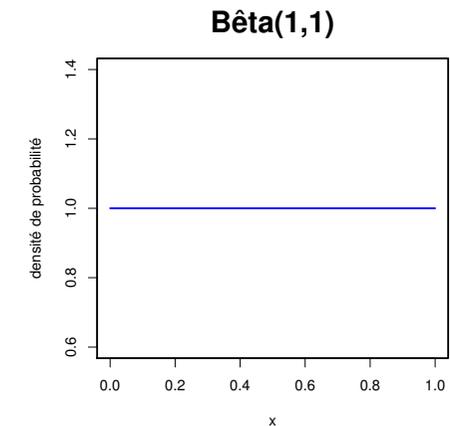
$$\hat{p} = \frac{r}{n} \quad ?$$



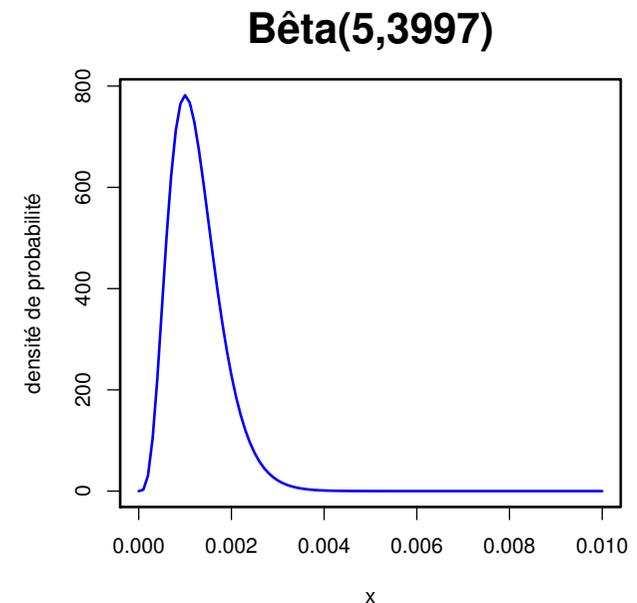
## Ex. approche classique

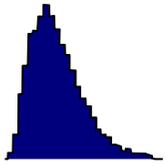
1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Inférence Bayésienne avec loi *a priori* non informative :  $\text{Bêta}(1,1)$



→ Loi *a posteriori* pour  $p$  :  
 $\text{Bêta}(r+1, n-r+1) =$   
 $\text{Bêta}(5, 3997)$





## Ex. approche classique modifiée

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

- Inférence Bayésienne avec loi *a priori*

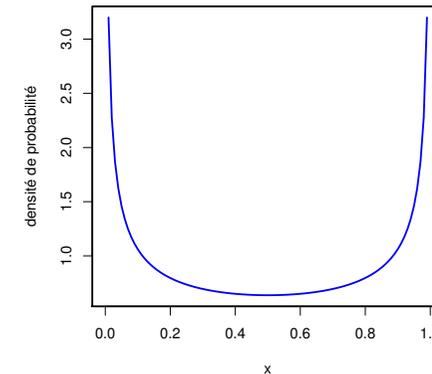
Bêta(0.5,0.5)

(parfois utilisée pour ses propriétés mathématiques intéressantes)

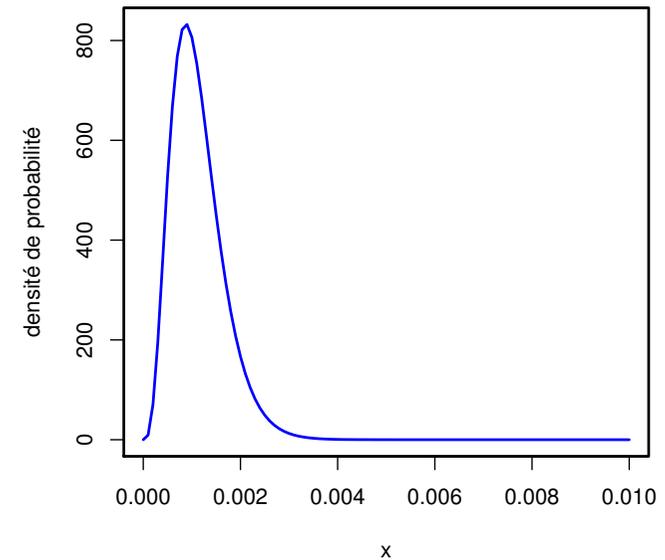
→ Loi *a posteriori* pour  $p$  :

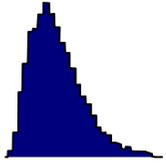
$$\text{Bêta}(r + 0.5, n - r + 0.5) = \text{Bêta}(4.5, 3996.5)$$

Bêta(0.5,0.5)



Bêta(4.5,3996.5)

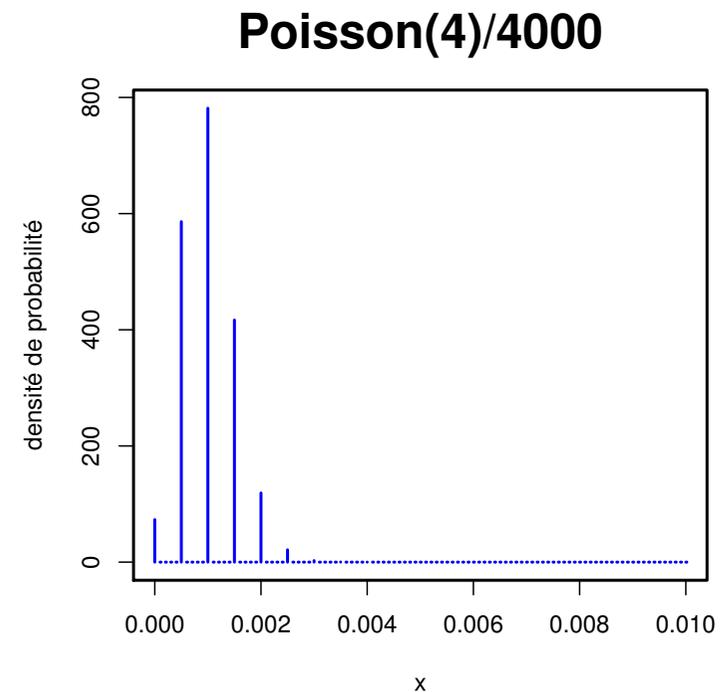


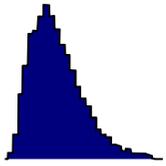


## Ex. bootstrap sur les 4000 lots

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

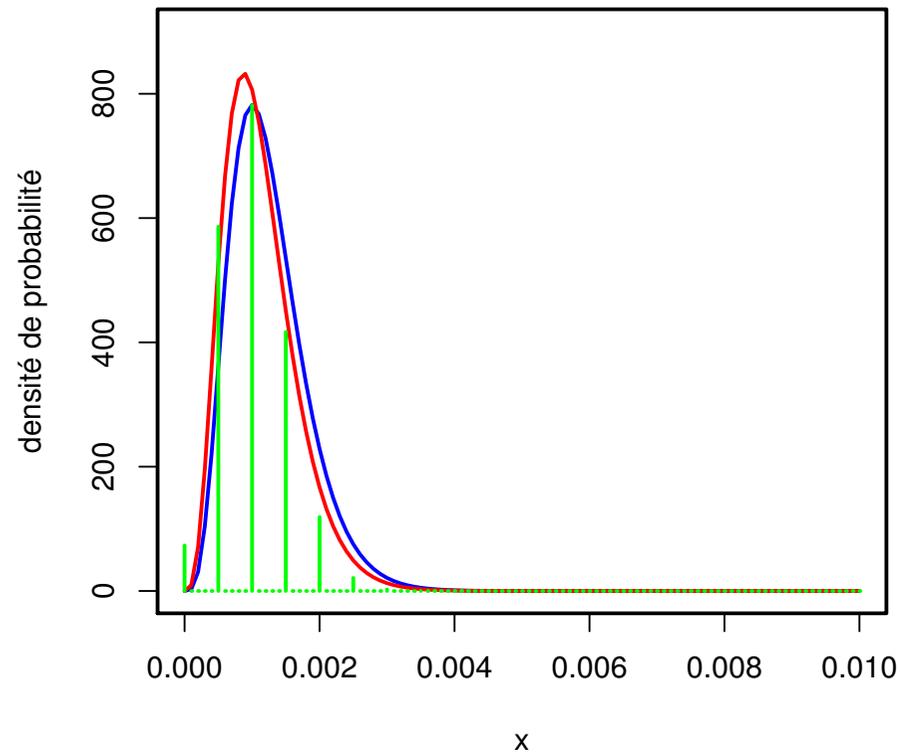
- Tirage au sort avec remise de 4000 lots dans les 4000 lots dont 4 sont positifs
- Loi suivie par l'estimation de  $p$  :  
 $\text{Binom}(4000, 0.001)/4000$   
 $\rightarrow \text{Poisson}(4)/4000$



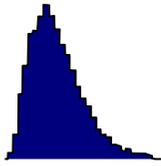


## Ex. comparaison des 3 lois d'incertitude

1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude



A titre indicatif, l'intervalle de confiance à 95% autour de la prévalence est  $[0.00027; 0.0026]$



## A retenir



1. Processus stochastiques
2. Nombreuses données
3. Dires d'expert
4. Lois d'incertitude

## Trois approches de modélisation de l'incertitude

### ■ Statistique fréquentiste

utilisation des intervalles de confiance

**Pas trivial !**

### ■ Statistique bayésienne

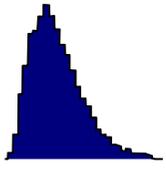
loi *a posteriori* sur un paramètre

→ loi d'incertitude

### ■ Rééchantillonnage (bootstrap)

échantillon bootstrap d'un paramètre

→ loi d'incertitude empirique



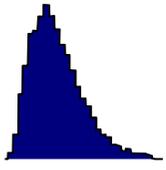
## Conclusion

### **Pour spécifier une loi de variabilité ou d'incertitude il est important**

- de connaître les lois classiques et leur cadre d'utilisation
- d'utiliser une méthodologie rigoureuse

### **MAIS aussi**

- de toujours visualiser la loi obtenue
- de valider le choix avec le regard du biologiste



## Références bibliographiques

- CULLEN A.C. et FREY H.C. - Probabilistic techniques in exposure assessment. 335 pages, Plenum Press, New York, 1999.
- MORGAN M.G. et HENRION M. – Uncertainty. A guide to dealing with uncertainty in quantitative risk and policy analysis, 332 pages, Cambridge University Press, Cambridge, 1990.
- VOSE D. - Risk analysis. A quantitative guide. 418 pages, Wiley. Chichester, 2000.