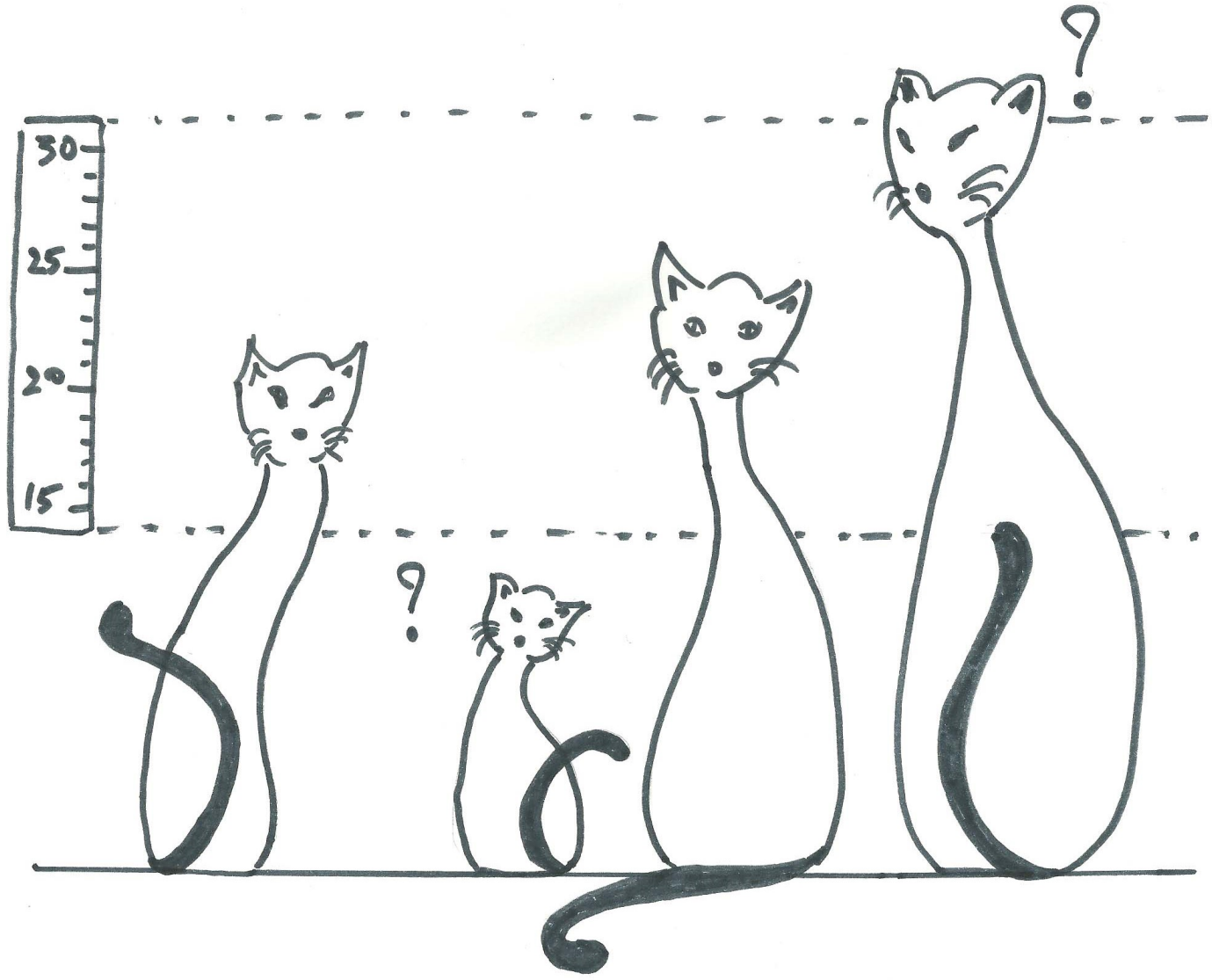
Que faire quand certaines données sont en dessous de la limite de quantification? Introduction à la problématique des données censurées à gauche

M.L. Delignette-Muller

22 novembre, 2018

Qu'est-ce qu'une donnée censurée ?



Définition des types de censures

On parle de données censurées lorsque pour certaines observations d'une variable quantitative x on ne dispose pas d'une valeur numérique mais d'une information de type $x > a$ ou $x < a$ ou $a \leq x \leq b$.

- **censure à gauche :** $x < a$, avec a une limite de quantification par ex.
- **censure à droite :** $x > a$, avec a le temps au bout duquel un patient est sorti vivant d'une étude de survie par ex.
- **censure par intervalle :** $a \leq x \leq b$, avec a et b les deux temps d'observations entre lesquels on sait qu'un organisme est mort dans une étude de survie par ex.

De nombreuses méthodes exsitent pour les censures à droite dans le cadre de l'étude de la survie (présentation de Karine à venir).

Nous allons ici nous focaliser sur les censures à gauche.

Un exemple en microbiologie

Estimation de la concentration en bactéries d'une espèce pathogène donnée dans une matrice alimentaire.

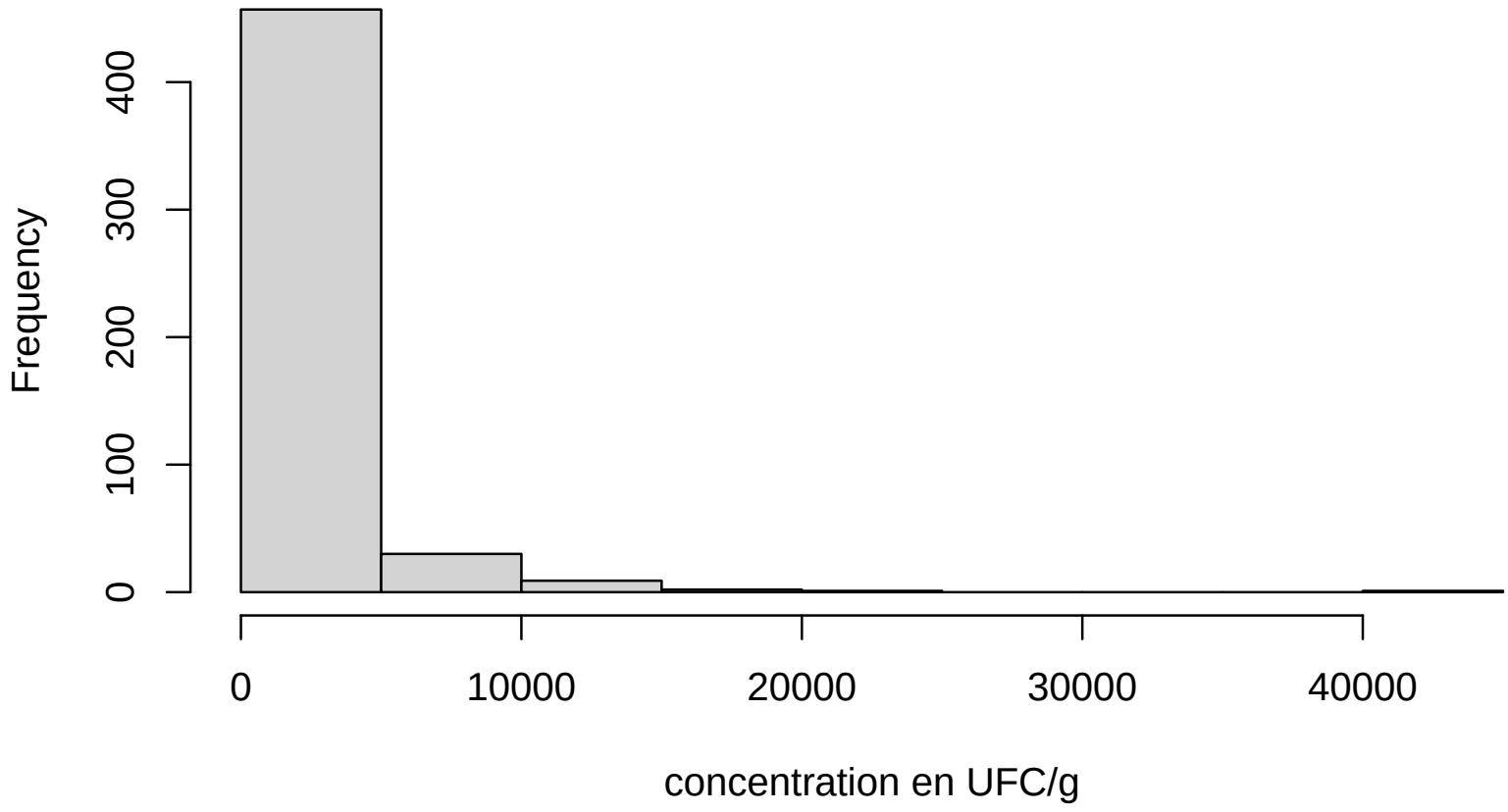
Un protocole classique :

- dilution de 25 g de matrice alimentaire solide dans 225 ml de bouillon de culture
- homogénéisation du mélange
- étalement d'un échantillon de 0.1 ml du mélange sur une boîte de Petri
- incubation à température suboptimale
- comptage des colonies

Concentration estimée en UFC.g^{-1}.

Plus petite concentration estimée non nulle : une colonie pour un équivalent de 0.01 g de matrice étalée = 100 UFC.g^{-1}.

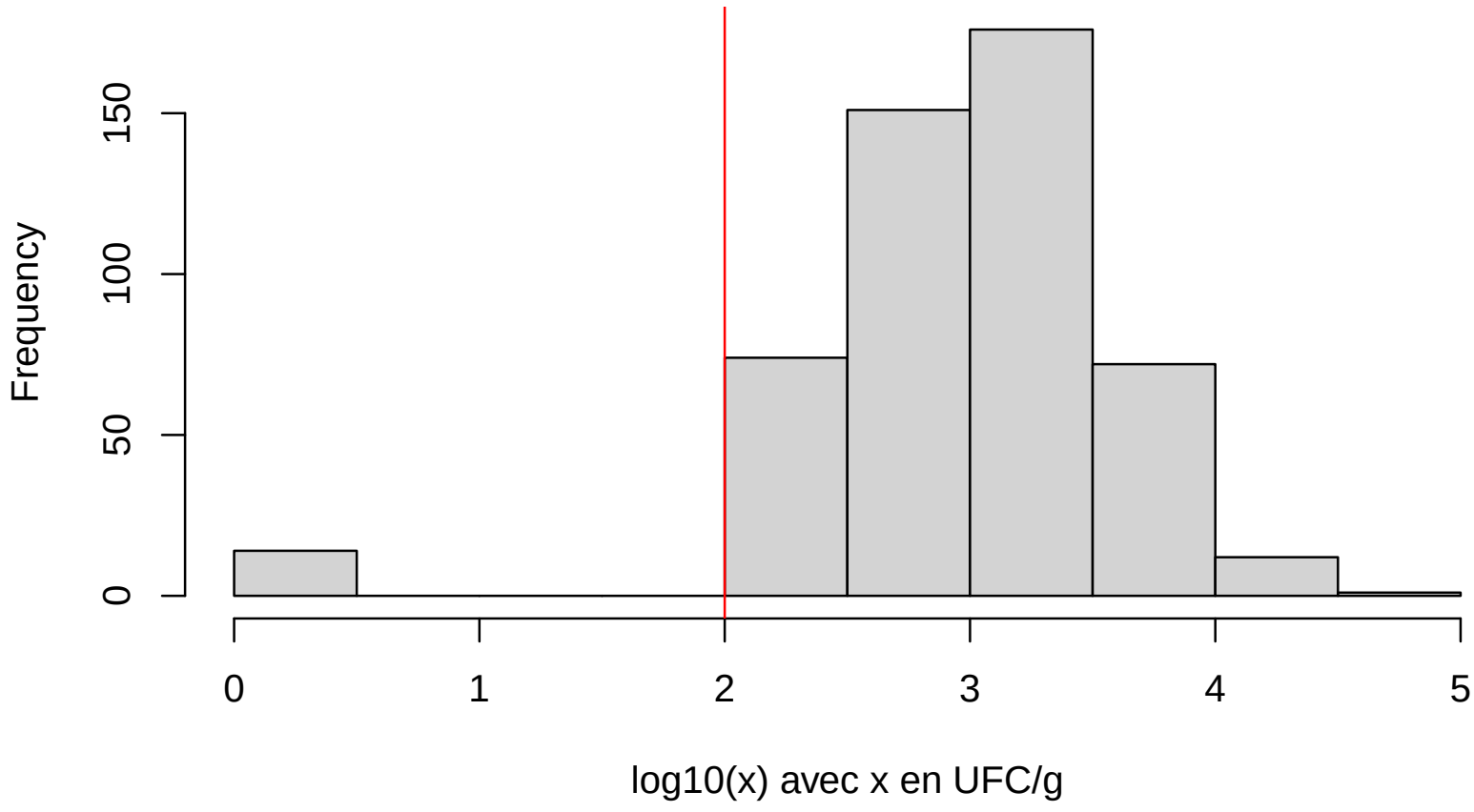
Jeu de données - histogramme



Transformation des données pour l'analyse statistique

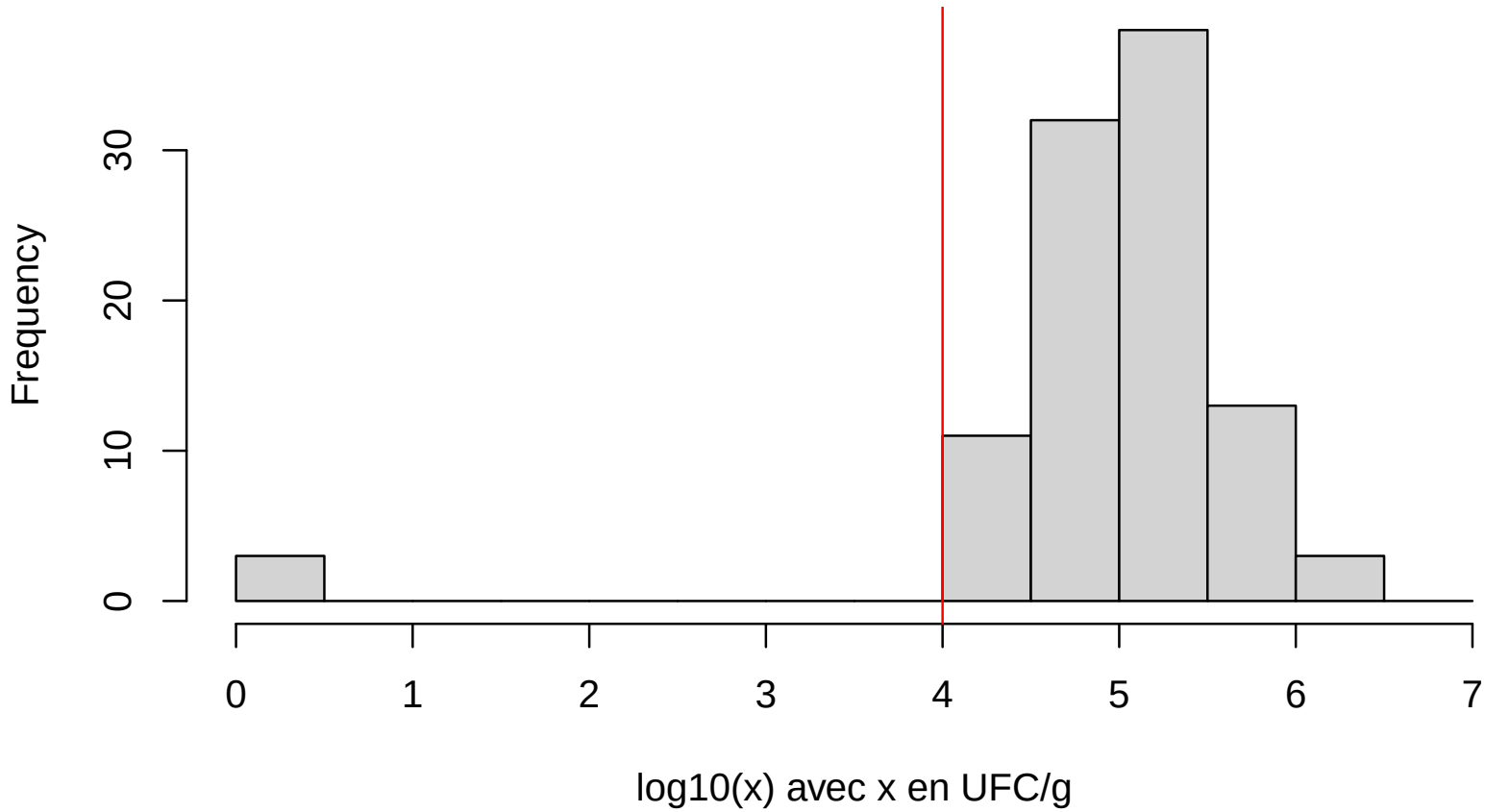
- Pour **normaliser la distribution** on serait tenté d'utiliser une **transformation logarithmique**, mais cela est **impossible du fait de la présence de zéros**.
- Les zéros correspondent en fait à des **données censurées**. On sait juste que la concentration correspondante est inférieure à la limite de sensibilité (ici 100 UFC.g^{-1}).
- Certains vont contourner le problème en remplaçant les zéros par des 1 ou en appliquant la transformation $log_{10}(x+1)$ (approche assez courante semble-t-il).
- Mieux vaut ne pas modifier les données non censurées (donc remplacer uniquement les 0) que de transformer toutes les données avec $log_{10}(x+1)$ mais pourquoi remplacer les zéros par des 1 (valeur arbitraire)?

Jeu de données avec les 0 transformés en 1



Dans cet exemple cela met les données censurées sans doute un peu trop loin du reste de la distribution.

Jeu de données avec les 0 transformés en 1 si on multiplie x et sa limite de quantification par 100



La valeur arbitraire serait encore plus décalée de la vraie distribution.

Par quelle valeur non arbitraire remplacer les données censurées lorsqu'elles sont en petit nombre ?

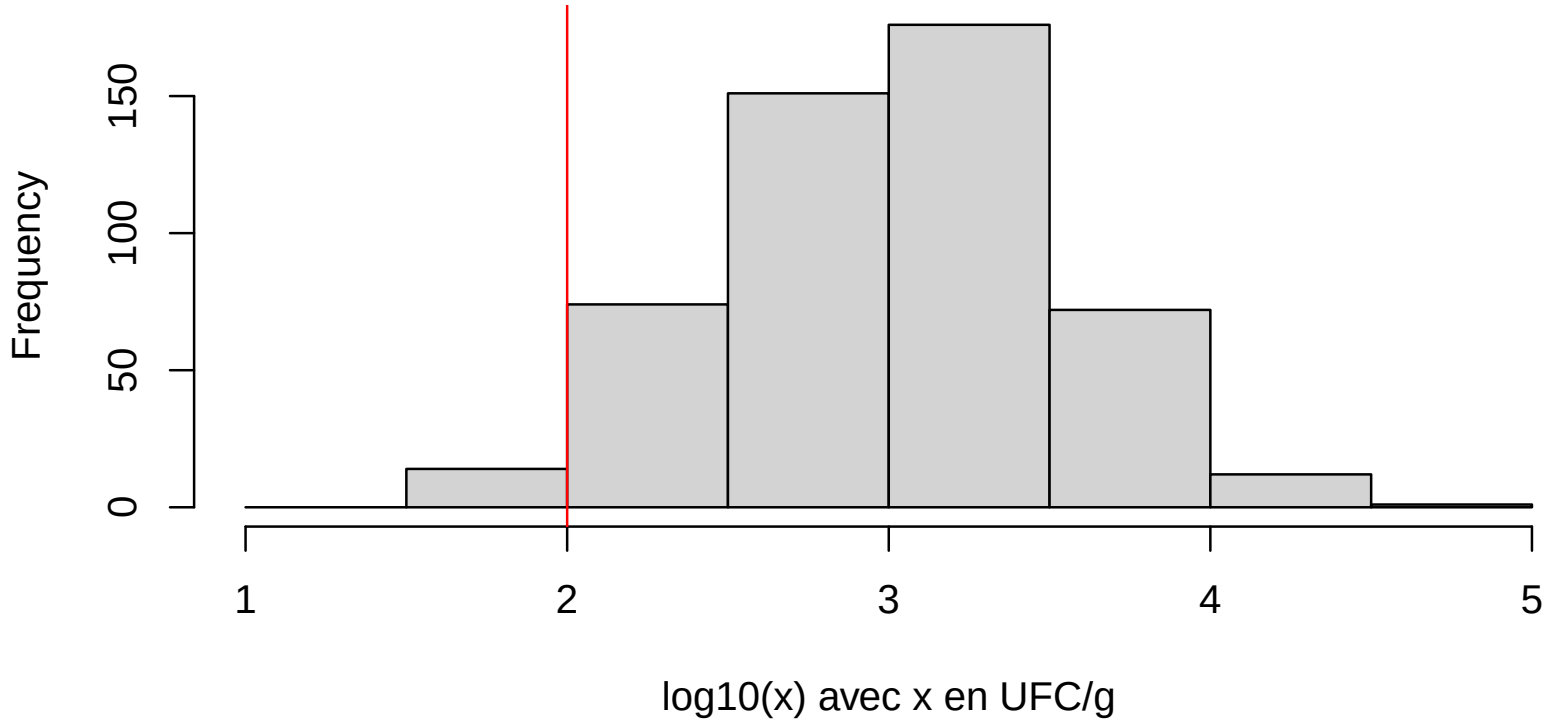
Il ne semble pas idiot de tenter de remplacer les données censurées par une valeur non nulle si les données censurées sont en petit nombre, mais il convient de choisir intelligemment la valeur de remplacer.

Il semple utile de regarder les plus petites concentrations estimées non nulles et la limite de quantification:

- 3 colonies : 300 UFC.g^{-1}.
- 2 colonies : 200 UFC.g^{-1}.
- 1 colonie : 100 UFC.g^{-1}.
- 0 colonie : ?

Au vu des valeurs précédentes il ne serait pas idiot de définir la valeur de remplacement à quelques dizaines d'UFC.g^{-1}., par exemple 50.

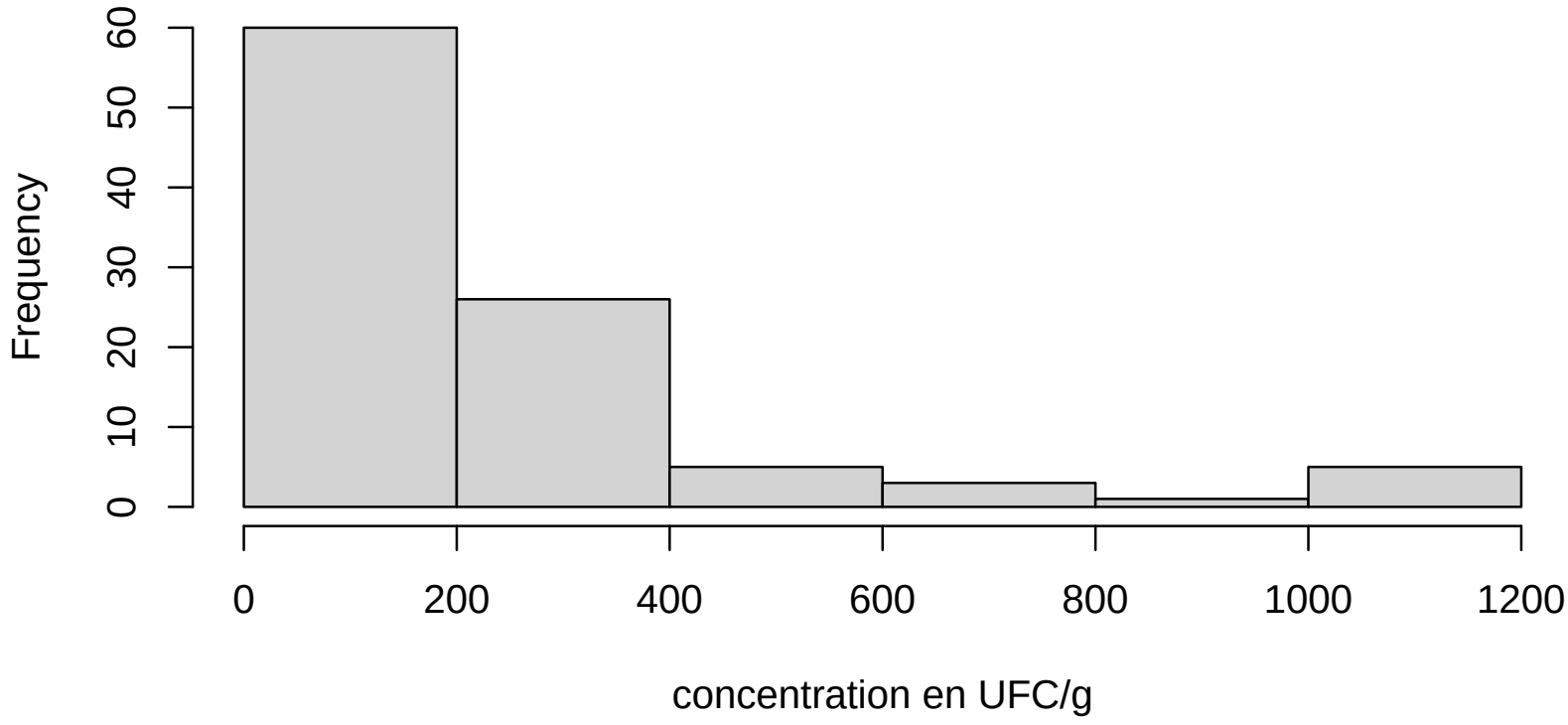
Jeu de données initial avec les 0 remplacés ici par 50 $UFC.g^{-1}$.



On s'approche sans doute plus de la vraie distribution (celle qu'on aurait s'il n'y avait pas de limite de quantification).

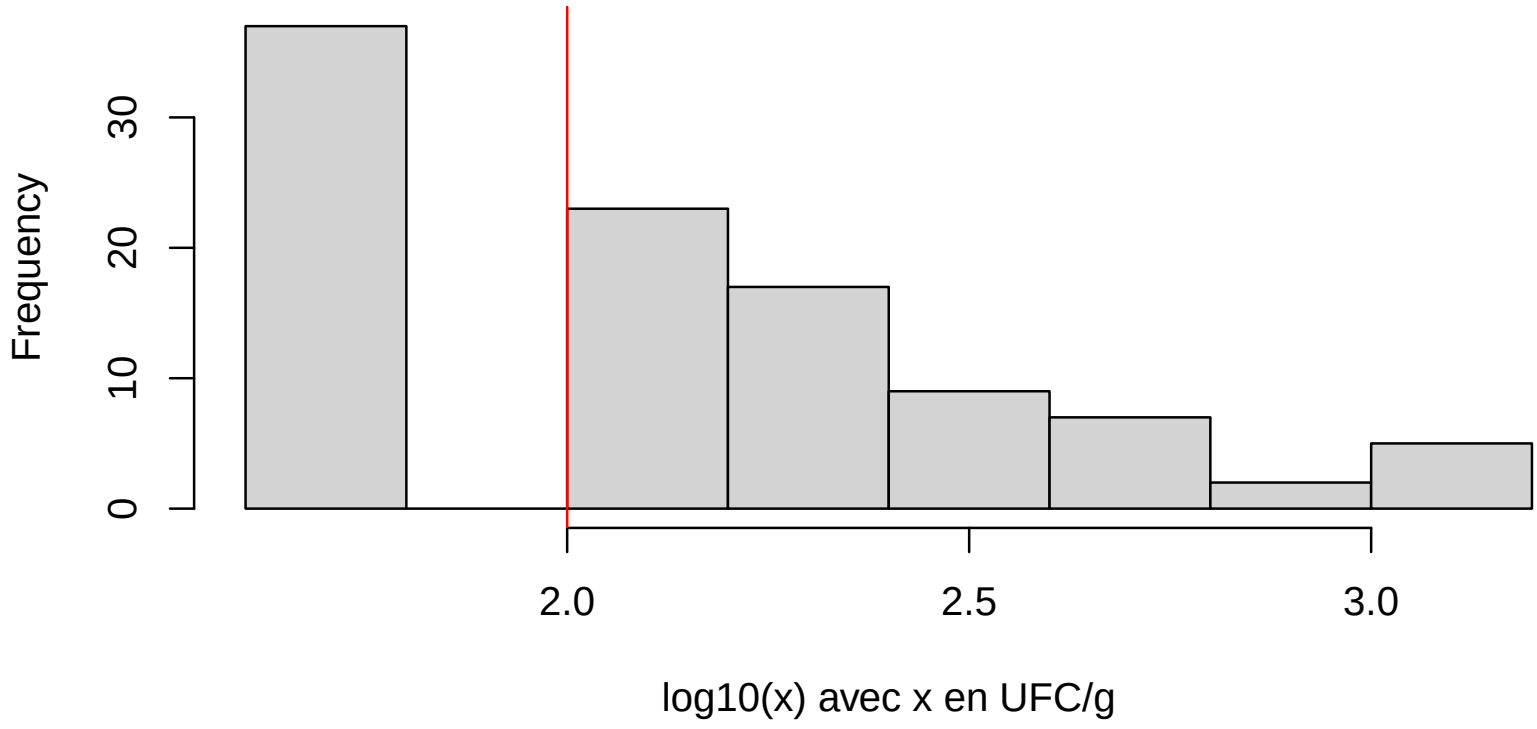
Que se passe-t-il si le jeu le nombre de censures est plus élevé ?

Imaginons que le niveau de x soit divisé par 10.



La proportion de valeurs censurées s'approche de 50%.

Jeu de données initial avec les 0 remplacés ici par 50 $UFC.g^{-1}$.



Quelle que soit la valeur par laquelle on va remplacer les 0, on n'arrivera jamais à normaliser la distribution. Que faire alors ?

Solutions possibles dans le cas d'un grand nombre de données censurées

Solutions simples mais grossières

- utilisation de méthodes non paramétriques (tests des rangs)
- transformation de x en variable qualitative (contaminée ou non)

Solutions plus sophistiquées

- mise en oeuvre de méthodes **prenant en compte la censure** telle que, par ex. par **maximum de vraisemblance** (ex. package NADA pour représenter et comparer des données censurées à gauche ou package fitdistrplus pour ajuster une distribution paramétrique sur données censurées)
- **modélisation des données brutes** (ex.: nombre de colonies observées sur la boîte de Petri) en décrivant tous les processus stochastiques et **inférence bayésienne**.

Un exemple de codage de données censurées à gauche pour utilisation avec NADA

```
str(d)
```

```
## 'data.frame':    200 obs. of  3 variables:
##  $ obs  : num  300 900 11000 8100 300 200 600 100 1600 1
##  $ cen  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ group: Factor w/ 2 levels "G1","G2": 1 1 1 1 1 1 1 1
```

Visualisation des données

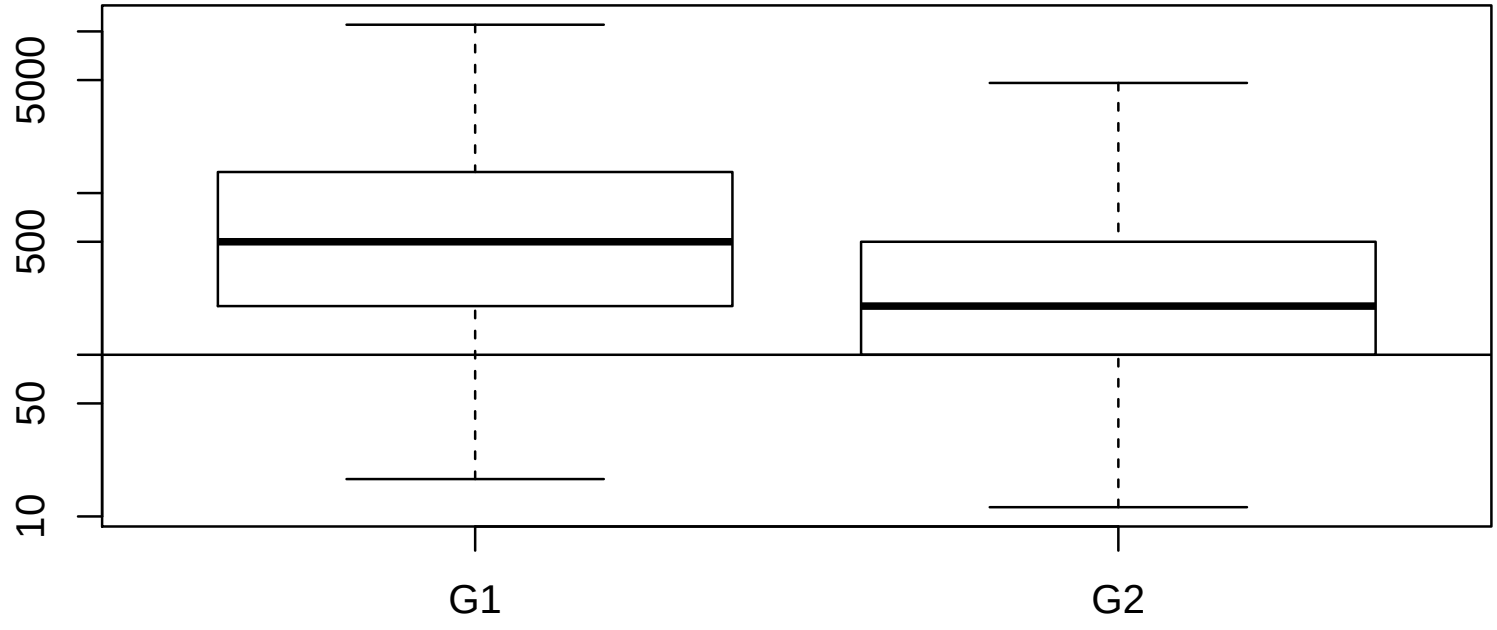
Les valeurs censurées son fixées à la limite de quantification dans la variable obs et la censure est indiquée dans la variable logique cen.

```
d
```

```
##        obs   cen group
## 1      300 FALSE    G1
## 2      900 FALSE    G1
## 3    11000 FALSE    G1
## 4     8100 FALSE    G1
## 5      300 FALSE    G1
## 6      200 FALSE    G1
## 7      600 FALSE    G1
## 8      100 FALSE    G1
## 9     1600 FALSE    G1
## 10    1200 FALSE    G1
## 11     300 FALSE    G1
## 12     100 FALSE    G1
## 13     100  TRUE    G1
```

Un exemple de représentation de diagrammes en boîte sur données censurées à gauche avec NADA - attention, ne pas trop se fier à ce qui est en dessous de la limite !

```
require(NADA)
cenboxplot(d$obs, d$cen, d$group, log = TRUE)
```



Un exemple de codage de données censurées à gauche pour utilisation avec fitdistrplus

```
str(dgroup2)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ left : num  200 1300 200 500 400 100 500 100 200 1000
##  $ right: num  200 1300 200 500 400 100 500 100 200 1000
```

Visualisation des données

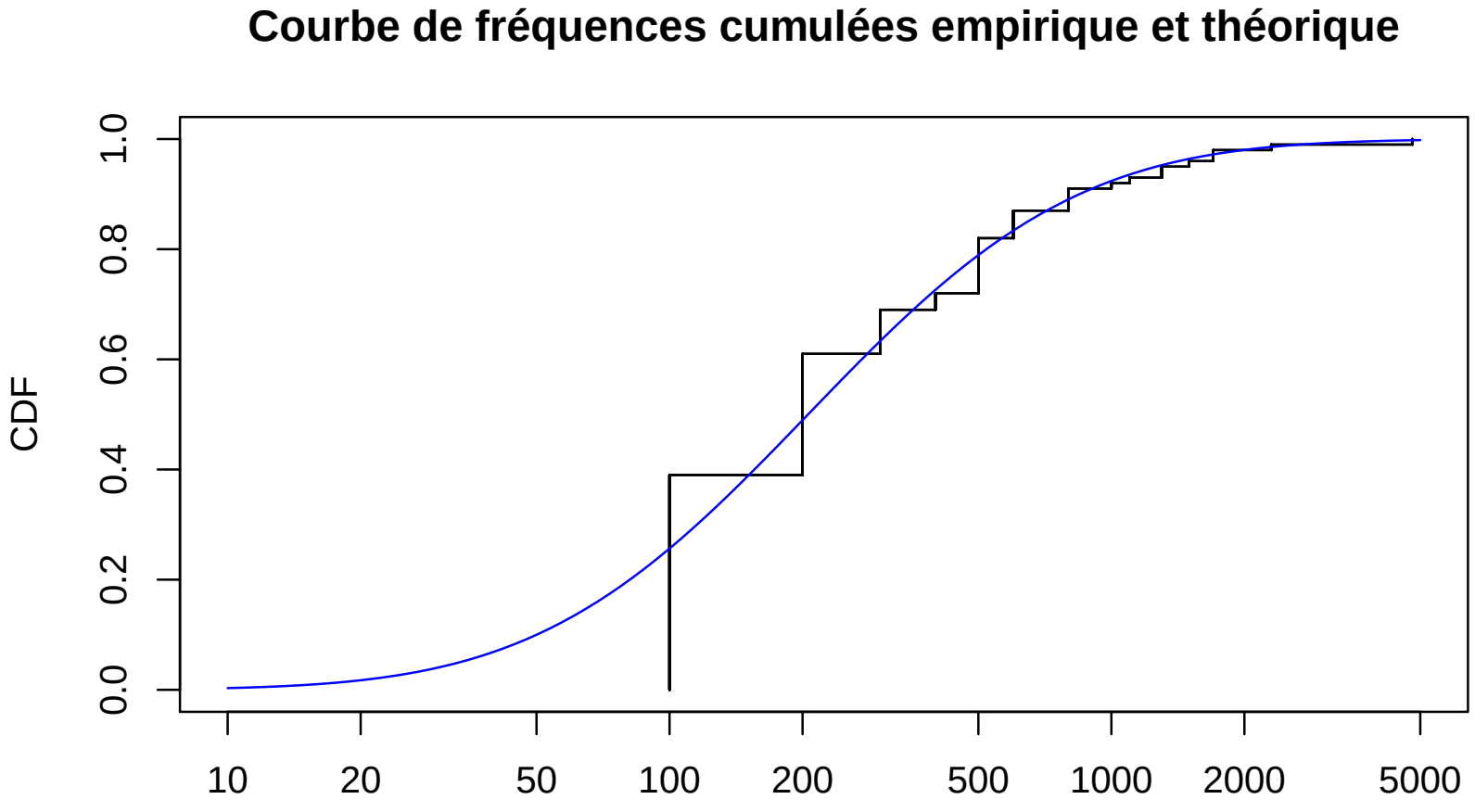
Les observations sont codées avec deux valeurs, left et right, toutes deux égales à la valeurs en absence de censure, et left à NA et right au seuil de censure en cas de censure à gauche

```
dgroup2
```

```
##      left right
## 1     200   200
## 2    1300  1300
## 3     200   200
## 4     500   500
## 5     400   400
## 6     100   100
## 7     500   500
## 8     100   100
## 9     200   200
## 10   1000  1000
## 11     NA   100
```

Un exemple de représentation d'ajustement d'une distribution lognormale avec fitdistrplus

```
require(fitdistrplus)
fit <- fitdistcens(dgroup2, "lnorm")
cdfcompcens(fit, xlogscale = TRUE, xlim = c(10, 5000), fitc
```



Une référence pour aller plus loin

Helsel, D. R. (2005). Nondetects and data analysis. Statistics for censored environmental data. Wiley-Interscience.

Le package NADA est associé à cet ouvrage.