ATELIER Proba / Stat (Partie 2)
vendredi 15 / 06 / 2012

Calculs de fractiles empiriques à l'aide des logiciels et calculatrices

P. GARAT [(1)], Florent GIROD[(2)]

(1) Université de Grenoble, Dept. STID
(2) IREM de Grenoble

PLAN

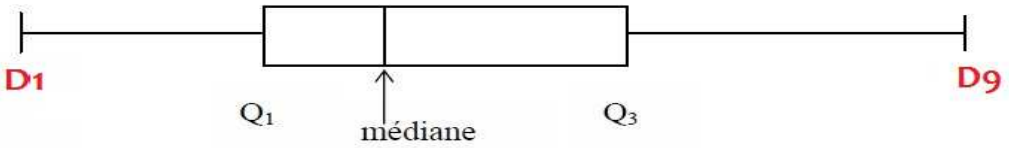
1. Introduction

Importances des fractiles dans la décision statistique

- Les fractiles d'une distribution permettent de construire aisément des **intervalles de fluctuation** pour une variable étudiée, et de façon corrolaire des **régions critiques**, des **seuils critiques** ;
- On utilise les fractiles de la loi binomiale dans **le test exact d'une proportion** ;
- Le **rôle de la médiane** est essentiel en tant que paramètre de localisation ;
- De nombreux ouvrages en Génie Civil sont dimensionnés par rapport aux **valeurs centennales** de variables climatiques (hauteur d'eau, neige, force du vent, …)
- Autre exemple en santé publique : analyse de sang, ...

Les fractiles dans l'enseignement secondaire

- Les fractiles usuels **médiane et quartiles** sont enseignés dès la classe de Seconde ;
- Le diagramme en boîte est enseigné en Première. Il est parfois préconnisé d'utiliser les déciles D1 et D9 pour tronquer les « moustaches » du diagramme.



- En Terminale S : le seuil $u_{0.05}$ est en réalité le fractile de probabilité 97.5 %
- Des applications statistiques autour des fractiles sont possibles en Economie, SVT, Environnement, etc ...
- Le calcul des fractiles peuvent donner lieu à des exercices en Algorithmique.

2. Fractiles d'une loi de probabilité :

cas d'une loi discrète

Définition : *Fractiles d'une loi discrète*

Soient $a_1 < a_2 < \ldots < a_k < \ldots$ l'ensemble des modalités d'une variable discrète X
et soient $p_1, p_2, \ldots, p_k, \ldots$ les probabilités associées.
On appelle **fractile d'ordre *p*** de X (et on note x_p) la plus petite modalité a_j telle que la probabilité d'être inférieure ou égal à a_j soit au moins égal à p ; c'est-à-dire :

$$\text{Prob}[X < a_j] < p \quad \text{et} \quad \text{Prob}[X \leq a_j] \geq p \qquad (1)$$

Soit encore :

$$p_1 + p_2 + \ldots + p_{j-1} < p \quad \text{et} \quad p_1 + p_2 + \ldots + p_{j-1} + p_j \geq p$$

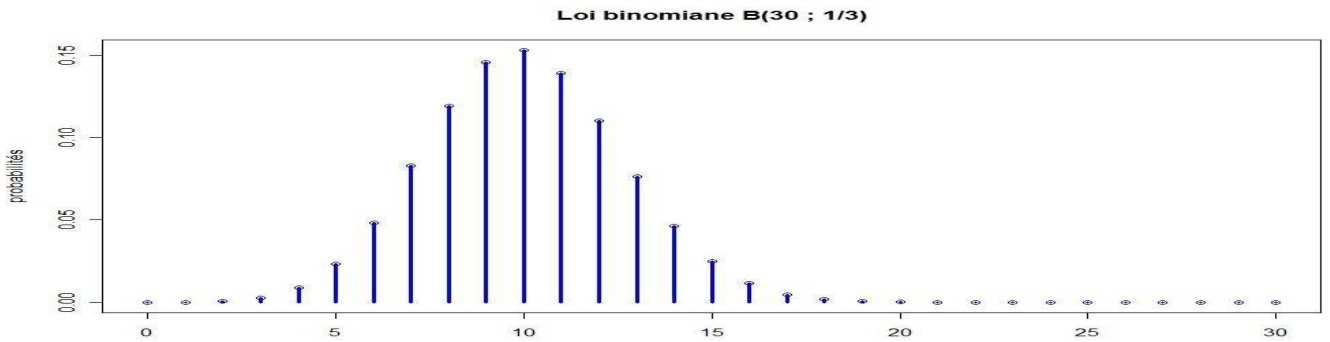
Considérant $F(x)$ la fonction de répartition de X, on peut aussi définir le fractile d'ordre p de X comme étant le plus petit élément de l'ensemble antécédent par F de $[\, p\, ;\, 1\,]$:

$$x_p = \min \{\, x \in \text{IR} \,/\, F(x) \geq p \,\}$$

Exemple

Considérons l'expérience aléatoire consistant à tirer **avec remise** 30 boules dans une urne contenant 1/3 de boules blanches.

Soit X = le nombre de boules blanches tirées.



Le tableau des probabilités cumulées donne :

a	0	1	2	3	4	5	6	7	8	9	10
Prob[$X \leq a$]	*0.000*	*0.000*	*0.001*	*0.003*	*0.012*	*0.035*	*0.084*	*0.167*	*0.286*	*0.432*	*0.585*
a	11	12	13	14	15	16	17	18	19	20	21
Prob[$X \leq a$]	*0.724*	*0.834*	*0.910*	*0.957*	*0.981*	*0.993*	*0.998*	*0.999*	*1*	*1*	*1*
a	22	23	24	25	26	27	28	29	30		
Prob[$X \leq a$]	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*		

Nous en déduisons par exemple **les fractiles** de la loi de X suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
5	6	7	8	10	12	13	14	15

Il est à noter que, par construction, la probabilité que *X* soit inférieur (strictement) à son fractile d'ordre *p* est au plus égale à *p*, et que la probabilité que *X* dépasse (strictement) son fractile d'ordre *p* est au plus égale à (1- *p*). Cette propriété sera très utile en statistique inférentielle pour la recherche de seuil critique. Nous l'écrivons :

$$\text{Prob}[X < x_p] < p \qquad \text{et} \qquad \text{Prob}[X > x_p] < 1 - p$$

Cela implique en particulier :

$$\text{Prob}[X < x_{p1} \ \ ou \ \ X > x_{1-p2}] < p_1 + p_2$$

On illustre **sur l'exemple :**

$$\text{Prob}[X < Q_{0.05} \ \ ou \ \ X > Q_{0.95}] < 10\ \%$$

Le calcul exact donne :

$$\text{Prob}[X < 6 \ \ ou \ \ X > 15] = 0.035 + 0.019 = 5.4\ \%$$

Fractiles empiriques d'une série statistique :

cas d'une variable statistique discrète

- Au-delà de la définition des fractiles théoriques d'une variable statistique X, il se pose le problème d'estimer ces fractiles à partir d'un échantillon $(x_1, x_2, \dots, x_n)$ **lorsque la loi de X est inconnue**.
  Nous appellerons fractiles empiriques de tels fractiles.

- Le cas discret ne pose *a priori* pas de problème particulier : il suffit de remplacer dans la définition (1) les probabilités théoriques p_1, p_2, …, p_k, … par les fréquences empiriques f_1, f_2, …, f_k, … obtenues lors de l'analyse fréquentielle (tri-à-plat) de l'échantillon (méthode 1).

- En pratique, deux autres méthodes sont parfois proposés selon les logiciels utilisés :

  ► **Trier l'échantillon $x_1, x_2, \dots, x_n$ par ordre croissant des valeurs ;**
  ► **Associer à chaque valeur x_j la *fréquence cumulée* $F(x_j) = j / n$**
  ► **Rechercher les deux valeurs d'indice *j-1* et *j* dont les fréquences cumulées encadrent la probabilité *p*, c'est-à-dire :**

  $$(j-1)/n < p \leq j/n$$

  ► **Calculer le fractile d'ordre *p* selon l'une des 3 méthodes :**

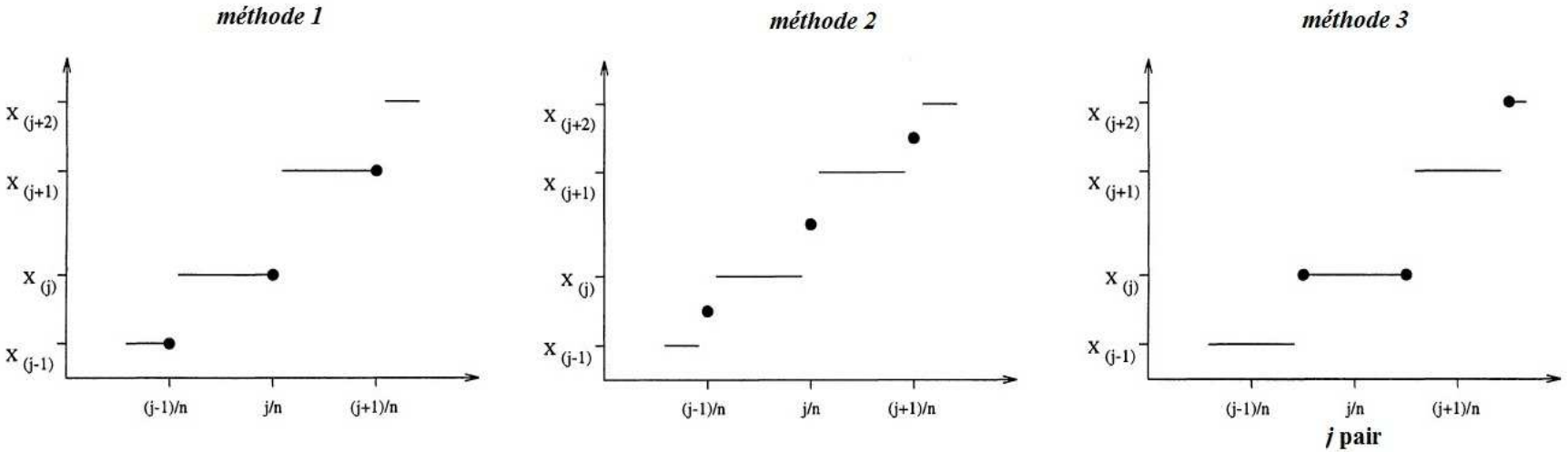
méthode 1 : $Q(p) = x_j$

méthode 2 : $Q(p) = x_j$ si $p < j/n$ et $Q(p) = (x_j + x_{j+1})/2$ si $p = j/n$

méthode 3 : $Q(p) = x_{j-1}$ ou x_j selon p est plus près de (j-1) ou de j
lorsque $p = j/n$ on choisit x_{j-1} ou x_j correspondant à un indice pair

Graphiquement :

Fonctions Quantile



Remarque :

La définition « lycée » de la médiane correspond à la méthode 2
La définition « lycée » des quartiles Q1 et Q3 correspond à la méthode 1

Exemples :

3. Expérimentations et Simulations

On souhaite savoir laquelle des trois méthodes donne de meilleurs résultats pour estimer les fractiles d'une loi binomiale B(30 ; 1/3) à partir d'un échantillon de taille n_{ech} = 20 ou d'un échantillon de taille n_{ech} = 50.

Des simulations Monte-Carlo (R = 10.000) avec le logiciel R donnent les résultats suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
5	6	7	8	10	12	13	14	15

→ ***Taux de « réussite » à trouver la bonne valeur des fractiles (en %) :***

n_{ech} = 20	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***29.76***	***31.73***	***36.66***	***49.41***	***49.69***	***38.45***	***39.50***	***32.41***	***26.91***
Méthode 2	***29.76***	***21.18***	***25.57***	***32.76***	***37.76***	***29.21***	***21.77***	***17.14***	***26.91***
Méthode 3	***29.76***	***31.73***	***36.66***	***49.41***	***49.69***	***38.45***	***39.50***	***32.41***	***26.91***

n_{ech} = 50	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***40.66***	***53.57***	***53.05***	***64.64***	***70.85***	***59.74***	***56.59***	***46.20***	***40.46***
Méthode 2	***40.66***	***53.57***	***46.01***	***64.64***	***65.35***	***59.74***	***39.41***	***46.20***	***40.46***
Méthode 3	***38.58***	***39.91***	***53.05***	***69.56***	***70.85***	***59.74***	***56.59***	***46.20***	***40.46***

4. Fractiles d'une loi de probabilité :

cas d'une loi absolument continue

Définition : ***Fractiles d'une loi abs. continue (loi à densité)***

Soit X une variable aléatoire absolument continue admettant une densité de probabilité $f(x)$ et une fonction de répartition $F(x)$ sur IR. On appelle **fractile d'ordre *p*** la valeur x_p de X telle que la probabilité d'être inférieure ou égal à x_p vaut <u>exactement</u> p ; c'est-à-dire :

$$\text{Prob}[X \leq x_p] = p \qquad (2)$$

Soit :

$$F(x_p) = p$$

L'existence et l'unicité de x_p est garantie par le caractère continue et croissant de la fonction de répartition $F(x)$. Le fractile x_p est l'antécédent de p par la fonction $F(x)$.

$$x_p = F^{-1}(p) \qquad (3)$$

Exemple de loi continue :

Considérons la loi exponentielle "unilatérale" sur IR^+ d'espérance unité, c'est-à-dire la loi définie par la densité :

$$f(x) = \exp(-x) \qquad x \geq 0$$

La fonction de répartition associée à cette loi est :

$$F(x) = 1 - \exp(-x) \qquad x \geq 0$$

Le fractile x_p d'ordre p s'obtient en résolvant en x :

$$F(x) = p$$

On obtient :

$$x_p = -\ln(1-p)$$

On en déduit par exemple **les fractiles** de la loi de X suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
0,0253	0,0513	0,1054	0,2877	0,6931	1,3863	2,3026	2,9957	3,6889

Fractiles empiriques d'une série statistique :

cas d'une variable statistique continue

Le cas continu est plus difficile que le cas discret, et de nombreux algorithmes existent, donnant des résultats ayant des propriétés mathématiques sensiblement différentes. Par exemple, le tableur Excel adopte la méthode suivante :

- **Trier l'échantillon $x_1, x_2, \ldots, x_n$ par ordre croissant des valeurs**
- **Associer à chaque valeur x_i la *fréquence cumulée corrigée :***

$$F_n^*(x_j) = (j-1)/(n-1)$$

- **Rechercher les deux valeurs x_{j-1} et x_j dont les fréquences cumulées encadrent la probabilité p, c'est-à-dire : $F_n^*(x_{j-1}) < p \leq F_n^*(x_j)$**
- **Calculer le fractile d'ordre p à l'aide de la formule d'interpolation linéaire :**

$$\tilde{Q}_p = [\,(F_n^*(x_j) - p) * x_{j-1} + (p - F_n^*(x_{j-1})) * x_j\,] / (F_n^*(x_j) - F_n^*(x_{j-1}))$$

- Cette méthode consiste donc à approcher la fonction de répartition théorique de X par une fonction continue et linéaire par morceaux, notée $F_n^*(x)$ et à rechercher l'antécédent de p par $F_n^*(x)$.

- Compte tenu de la définition de $F_n^*(x_i) = (j-1)/(n-1)$ on voit que cette méthode fournit une valeur de médiane empirique égale à : $x_{[(n+1)/2]}$ si n est impair, et (x[n/2] + x[(n+1)/2]) / 2 si n est pair.

- Cette méthode est une alternative à la démarche « naturelle » consistant à interpoler linéairement les fréquences cumulées non corrigées $F_n(x_j) = j/n$ ou à interpoler linéairement les milieu des paliers de la fonction de répartition empirique, c'est-à-dire : $F_n(x_j) = (j-0.5)/n$.

Fractiles empiriques d'une série statistique continue :

Cinq méthodes disponibles :

Les diverses méthodes se différencient sur la manière de corriger les fréquences cumulées $F_n(x_j)$

méthode 4 : $F_n(x_j) = j / n$

méthode 5 : $F_n(x_j) = (j - 0.5) / n$

méthode 6: $F_n(x_j) = j /(n + 1)$

méthode 7 : $F_n(x_j) = (j - 1) / (n - 1)$ ← *EXCEL (et R par défaut)*

méthode 8 : $F_n(x_j) = (j - 1/3) / (n + 1/3)$

On justifie les méthodes 4 à 8 en remarquant que :

$$F(X_{(j)}) \text{ suit la loi Bêta } (j, n - j + 1)$$

Par conséquent : $E[F(X_j)] = j / (n + 1)$ et $\text{Mode}[F(X_j)] = (j - 1) / (n - 1)$

De manière approximative : $\text{Médiane}[F(X_j)] = (j - 1/3) / (n + 1/3)$

5. Expérimentations et Simulations

On souhaite savoir laquelle des cinq méthodes donne de meilleurs résultats pour estimer les fractiles d'une loi exponentielle Exp(1) à partir d'un échantillon de taille n_{ech} = 20 ou d'un échantillon de taille n_{ech} = 50.

Des simulations Monte-Carlo (R = 10.000) avec le logiciel R donnent les résultats suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
0,0253	0,0513	0,1054	0,2877	0,6931	1,3863	2,3026	2,9957	3,6889

→ ***Erreur relative absolue moyenne par rapport à la bonne valeur des fractiles (en %) :***

n_{ech} = 20	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***141***	***73***	***54***	***34***	***24***	***21***	***21***	***23***	***26***
Méthode 4	***142***	***72***	***53***	***34***	***25***	***21***	***21***	***24***	***25***
Méthode 5	***142***	***82***	***56***	***35***	***25***	***21***	***21***	***23***	***26***
Méthode 6	***142***	***71***	***52***	***35***	***25***	***22***	***25***	***31***	***26***
Méthode 7	***209***	***116***	***68***	***37***	***25***	***21***	***21***	***23***	***24***
Méthode 8	***142***	***74***	***53***	***35***	***25***	***21***	***22***	***25***	***26***

n_{ech} = 50	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***89***	***52***	***34***	***22***	***16***	***13***	***15***	***15***	***17***
Méthode 4	***60***	***45***	***34***	***22***	***16***	***13***	***14***	***15***	***17***
Méthode 5	***74***	***52***	***35***	***22***	***16***	***14***	***14***	***15***	***17***
Méthode 6	***60***	***45***	***34***	***22***	***16***	***14***	***15***	***17***	***24***
Méthode 7	***96***	***58***	***39***	***23***	***16***	***13***	***14***	***15***	***17***
Méthode 8	***67***	***49***	***35***	***22***	***16***	***14***	***14***	***16***	***19***

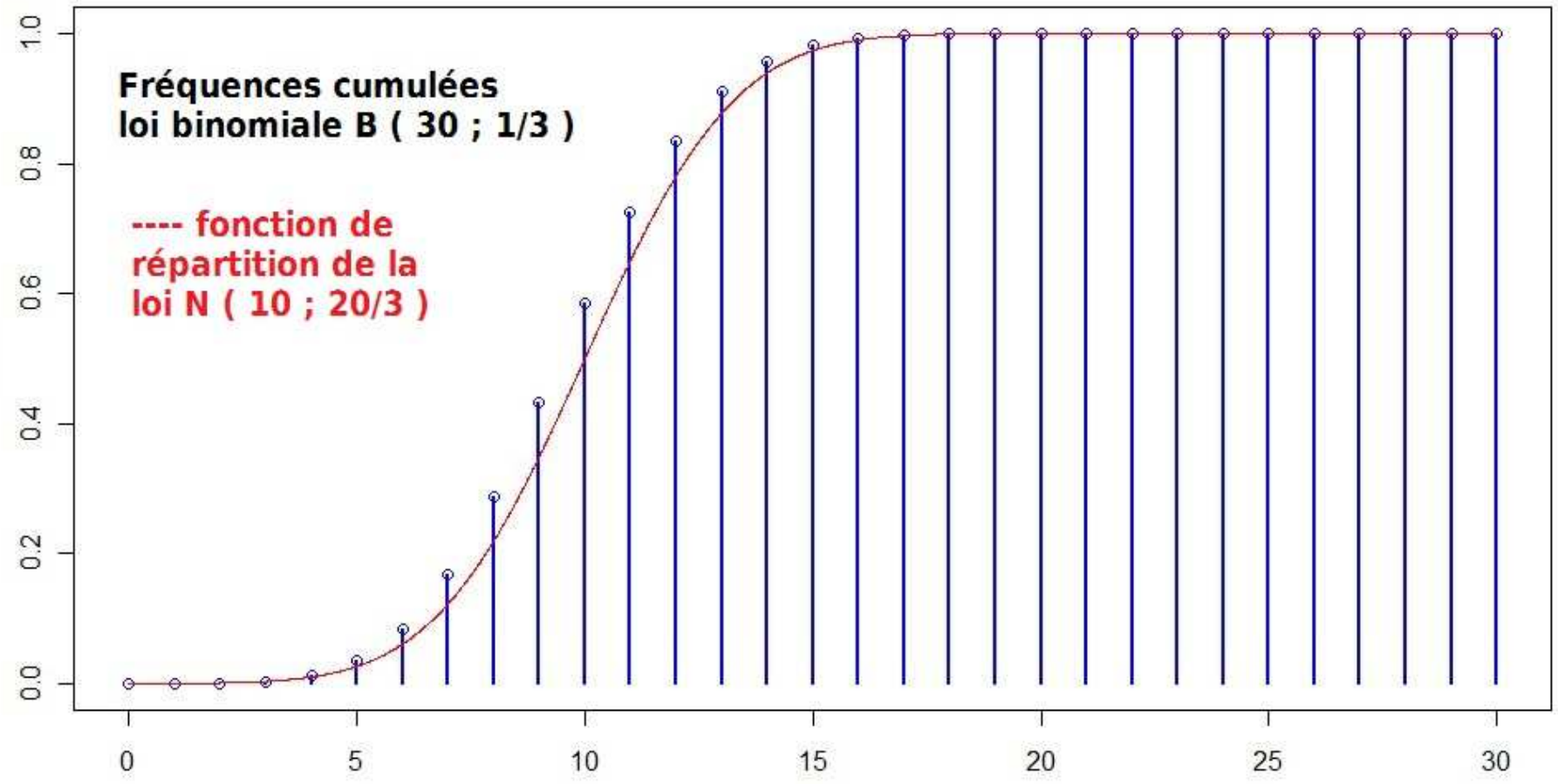
→ ***Biais d'estimation :***

$n_{ech} = 20$	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***0.02***	***0.00***	***0.00***	***-0.01***	***-0.02***	***-0.07***	***-0.22***	***-0.42***	***-0.11***
Méthode 4	***0.02***	***0.00***	***0.00***	***-0.01***	***-0.02***	***-0.07***	***-0.22***	***-0.42***	***-0.61***
Méthode 5	***0.02***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.08***	***-0.11***
Méthode 6	***0.02***	***0.00***	***0.00***	***0.01***	***0.03***	***0.08***	***0.23***	***0.53***	***-0.11***
Méthode 7	***0.05***	***0.05***	***0.05***	***0.04***	***0.03***	***-0.02***	***-0.17***	***-0.37***	***-0.59***
Méthode 8	***0.02***	***0.02***	***0.02***	***0.02***	***0.03***	***0.04***	***0.10***	***0.23***	***-0.11***

$n_{ech} = 50$	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	***0.01***	***0.01***	***0.00***	***0.01***	***-0.01***	***0.01***	***-0.09***	***0.00***	***-0.20***
Méthode 4	***0.00***	***0.00***	***0.00***	***0.00***	***-0.01***	***-0.03***	***-0.09***	***-0.17***	***-0.32***
Méthode 5	***0.01***	***0.01***	***0.01***	***0.01***	***0.01***	***0.01***	***0.01***	***0.00***	***0.05***
Méthode 6	***0.00***	***0.00***	***0.00***	***0.00***	***0.01***	***0.03***	***0.09***	***0.22***	***0.53***
Méthode 7	***0.02***	***0.02***	***0.02***	***0.02***	***0.01***	***-0.01***	***-0.07***	***-0.15***	***-0.31***
Méthode 8	***0.01***	***0.01***	***0.01***	***0.01***	***0.01***	***0.01***	***0.03***	***0.07***	***0.21***

6. Approximation des fractiles théoriques de la loi binomiale

Question : la correction du continu est-elle nécessaire ?



```
#-----------------------------------------------------------------
# approx des probas cumulées  de la loi binomiale
# binom(30,1/3)
#-----------------------------------------------------------------
n<-30
p<-1/3

pc.theo<- pbinom(0:n,n,p)

plot(0:n,pc.theo,col="blue",lwd=2,type="h")
points(0:n,pc.theo,col="blue")
x<-seq(0,n,by=0.01)
points(x,pnorm((x-n*p)/sqrt(n*p*(1-p))),col="red",type="l")
pc.approx1<-pnorm(((0:n)-n*p)/sqrt(n*p*(1-p)))
pc.approx2<-pnorm(((0:n)-n*p+0.5)/sqrt(n*p*(1-p)))
names(pc.theo)<-0:n
names(pc.approx1)<-0:n
names(pc.approx2)<-0:n

max(abs(pc.approx1-pc.theo))

0.0847596

max(abs(pc.approx2-pc.theo))

0.0085190
```

```
#-----------------------------------------------------------------------
# approx des fractiles de la loi binomiale
# binom(30,1/3)
#-----------------------------------------------------------------------
n<-30
p<-1/3
probs<-c(0.025,0.05,0.10,0.25,0.50,0.75,0.90,0.95,0.975)
q.theo<- qbinom(probs,n,p)
names(q.theo)<-probs

q.approx1<-n*p + qnorm(probs)*sqrt(n*p*(1-p))
q.approx1[probs<=0.5]<-ceiling(q.approx1[probs<=0.5])
q.approx1[probs>0.5]<-floor(q.approx1[probs>0.5])

q.approx2<-n*p + qnorm(probs)*sqrt(n*p*(1-p)) - 0.5*(probs<0.5) + 0.5*(probs>0.5)
q.approx2[probs<=0.5]<-ceiling(q.approx2[probs<=0.5])
q.approx2[probs>0.5]<-floor(q.approx2[probs>0.5])

names(q.theo)<-probs
names(q.approx1)<-probs
names(q.approx2)<-probs

q.theo
0.025  0.05   0.1  0.25   0.5  0.75   0.9  0.95 0.975
    5     6     7     8    10    12    13    14    15
q.approx1
0.025  0.05   0.1  0.25   0.5  0.75   0.9  0.95 0.975
    5     6     7     9    10    11    13    14    15
q.approx2
0.025  0.05   0.1  0.25   0.5  0.75   0.9  0.95 0.975
    5     6     7     8    10    12    13    14    15
```