Probabilités et Biostatistique

PCEM1 Pitié-Salpêtrière

A. Mallet et V. Morice

Cours 5 et 6

Statistiques descriptives

- Rappels et compléments (1)
  - variable aléatoire: caractéristique observable, variable, sans anticipation
  - expérience aléatoire: permet l'obtention d'une valeur (réalisation)
  - types usuels de variables aléatoires:
    - quantitatives: continues ou discrètes
    - qualitatives: nominales ou ordinales

Rappels et compléments (2)

- Définition de l'**unité statistique**:entité sur laquelle s'observe la variable aléatoire.
- Interprétation de la probabilité d'un événement aléatoire:
  - valeur limite de la fréquence avec laquelle l'événement se réalise au cours d'un nombre croissant de répétitions de l'expérience.

Rappels et compléments (3)

- la probabilité d’un événement aléatoire:
  - est une fiction
  - peut résulter d’une théorie
- pour l’approcher: répéter des expériences aléatoires
- les fréquences sont les contreparties expérimentales des probabilités

Echantillon (1)

- 1. Echantillon d’unités statistiques
  - n répétitions -> n unités statistiques concernées (sous population): constituent l’échantillon
- 2. Echantillon de valeurs de la variable étudiée X
  - tirage au ‘hasard pur’
  - i-ème répétition -> résultat x_i
  - $x_1, x_2, \ldots\ldots, x_n$: échantillon de valeurs de X

Echantillon (2)

- 3. Echantillon d’une variable aléatoire X
  - ensemble $X_1, X_2, \ldots\ldots\ldots X_n$ de n variables aléatoires, indépendantes entre elles, de même distribution que X
  - un échantillon de valeurs de X est une réalisation de l’échantillon de X
- n s’appelle la <u>taille</u> de l’échantillon

Echantillon (3)

X

x_1 x_2 x_3 x_4 premier échantillon de valeurs de X

x'_1 x'_2 x'_3 x'_4 second échantillon de valeurs de X

éalisations de X_1

éalisations de X_2

éalisations de X_3

éalisations de X_4

X_1,X_2,X_3,X_4 indépendantes entre elles

Echantillon (3)

X

x_1 x_2 x_3 x_4 premier échantillon de valeurs de X

première réalisation de X_1,X_2,X_3,X_4

x'_1 x'_2 x'_3 x'_4 second échantillon de valeurs de X

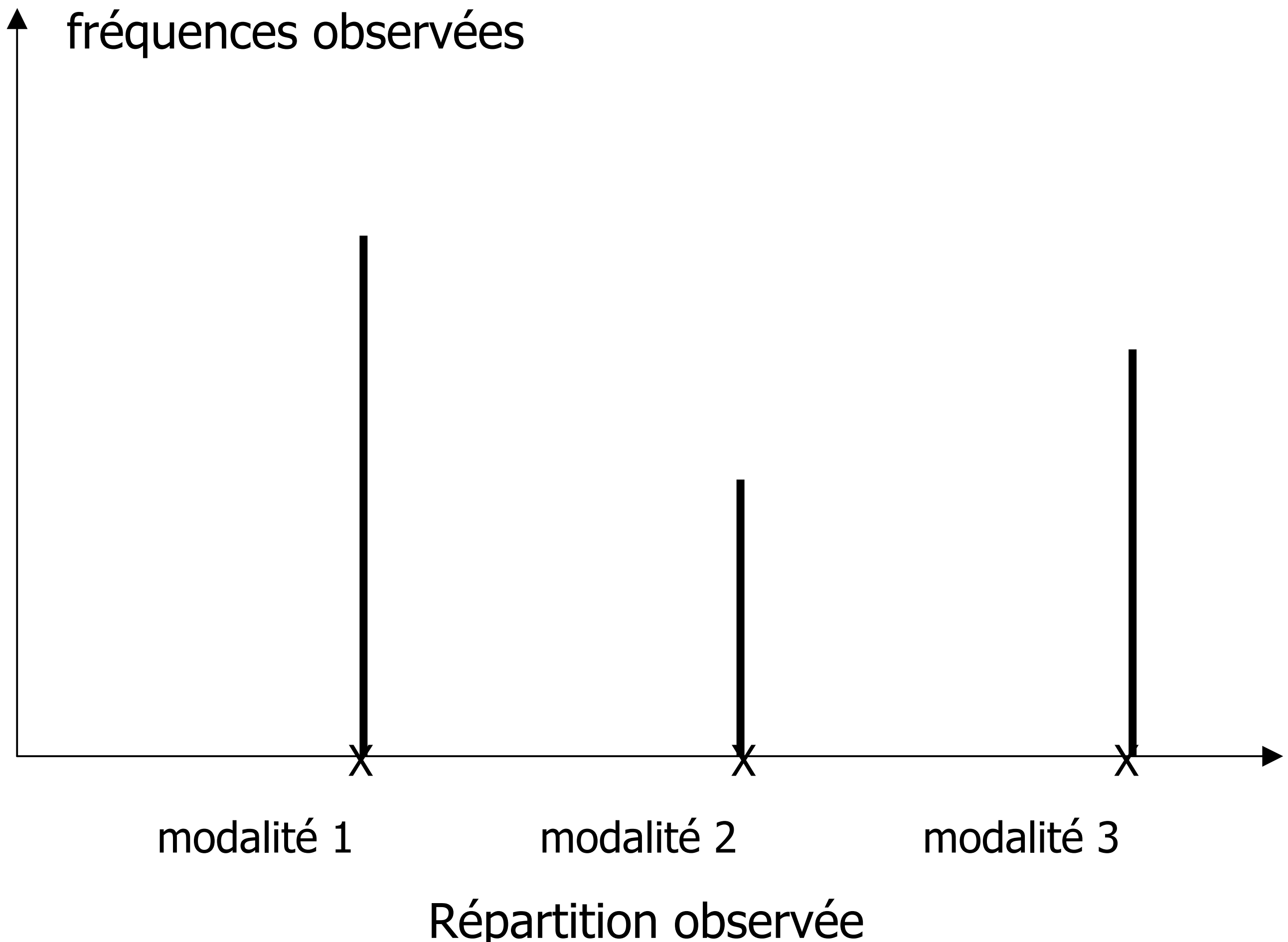
seconde réalisation de X_1,X_2,X_3,X_4

X_1 X_2 X_3 X_4 échantillon de la variable X

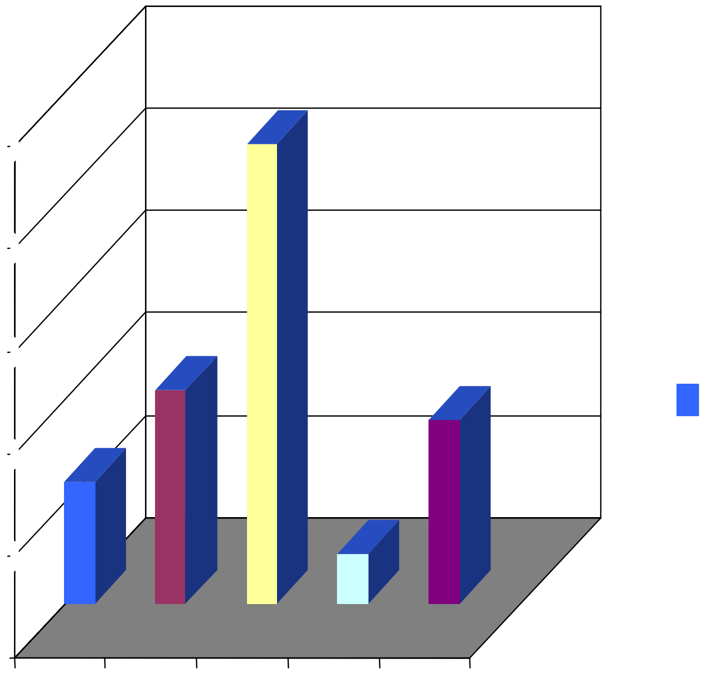
- un échantillon de valeurs de X est une réalisation de l'échantillon de X

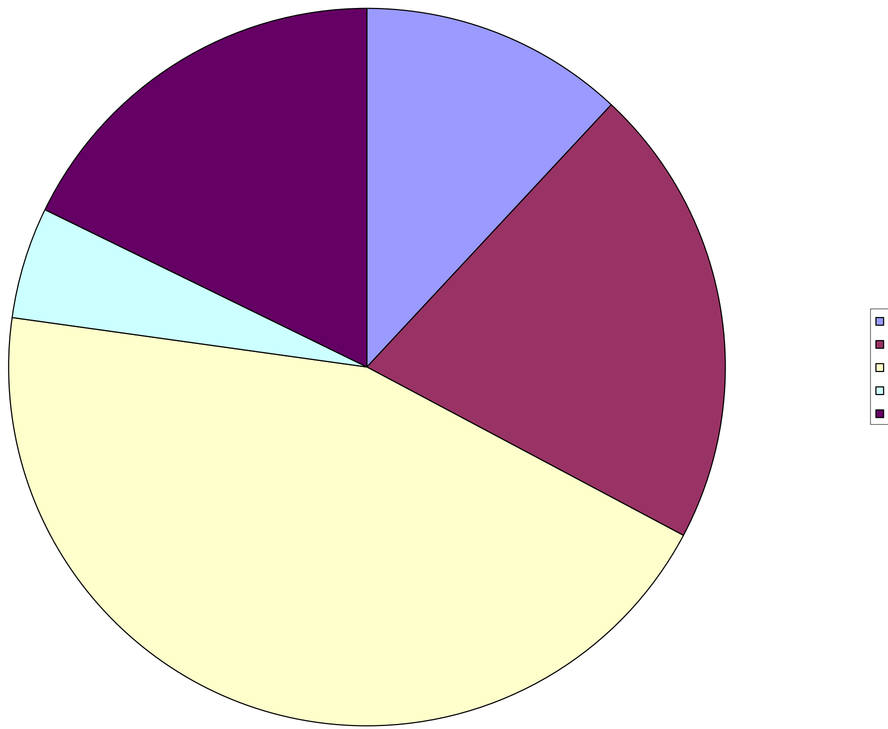
Représentation des résultats variable qualitative

- Référence: répartition



Représentation des résultats variable qualitative

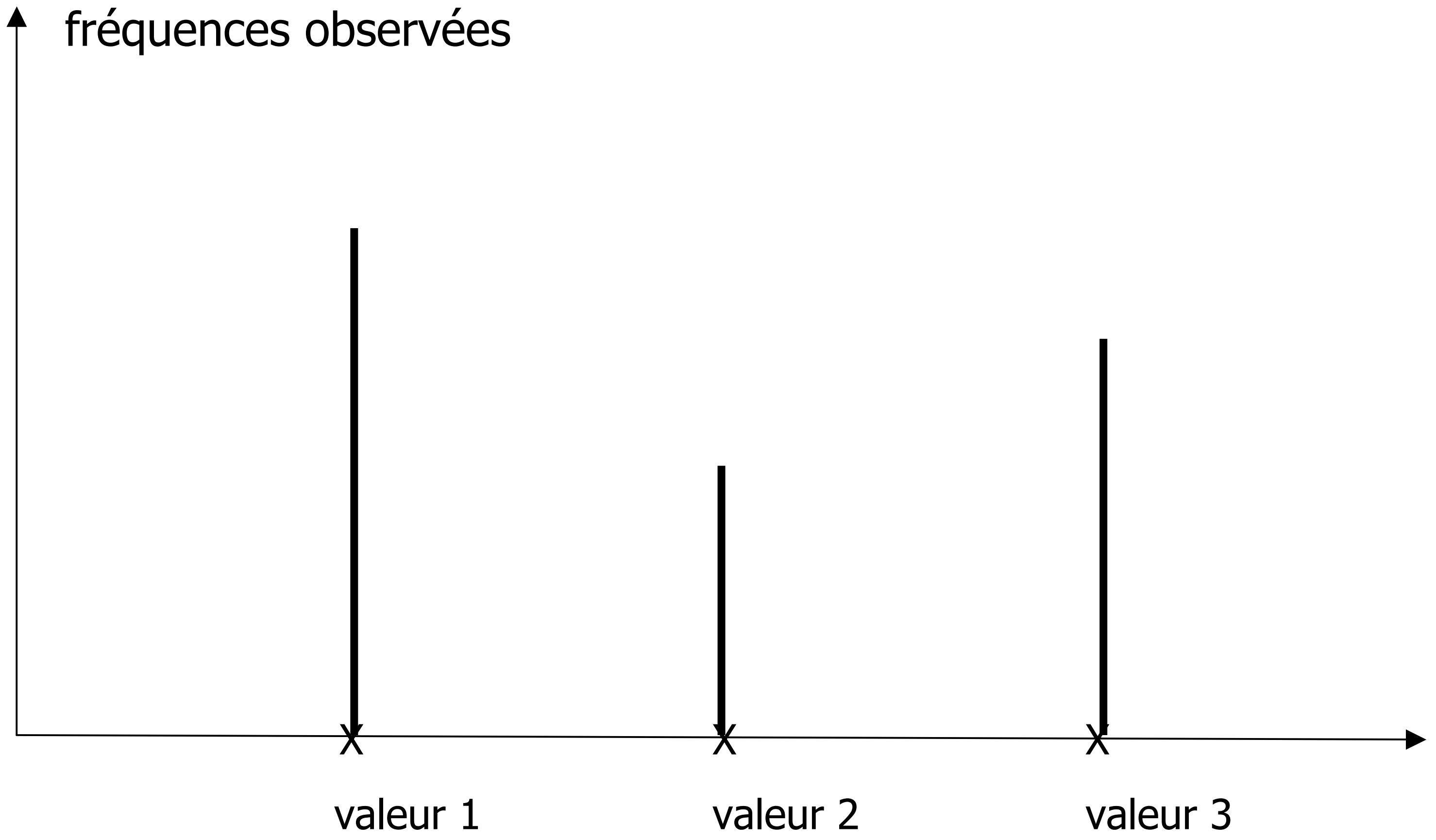




(variable à 5 modalités)

Représentation des résultats variable quantitative discrète

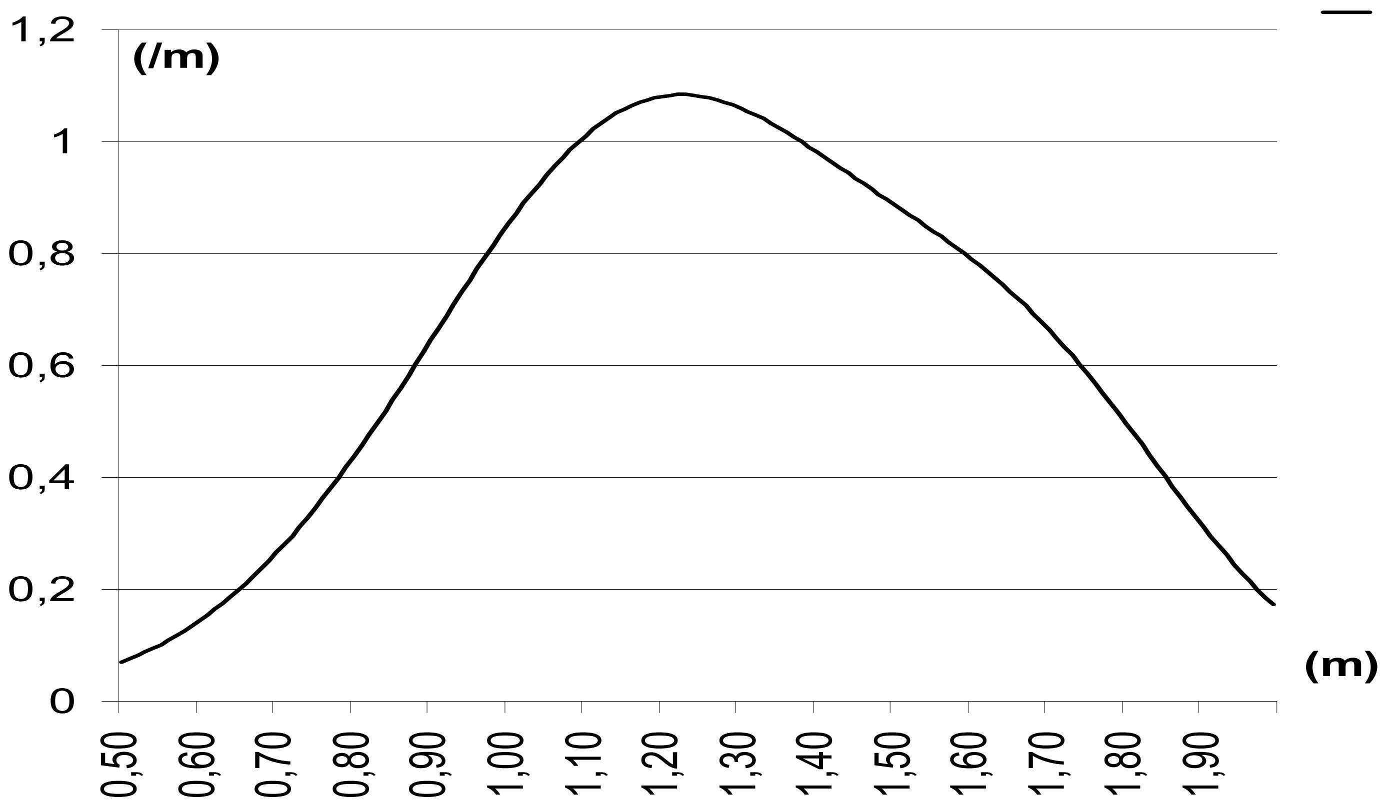
- Référence: répartition



Répartition observée ou histogramme en bâtons

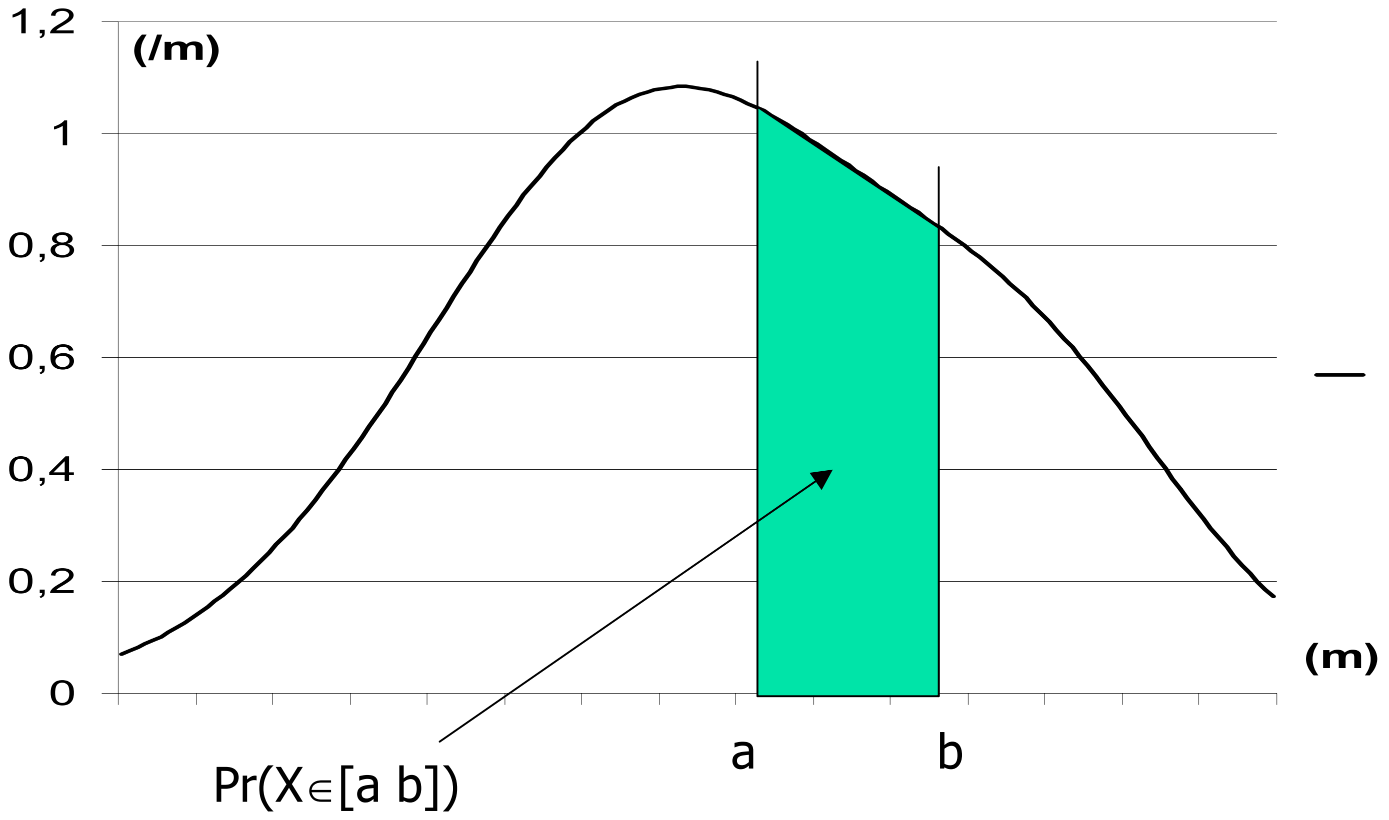
Représentation des résultats variable quantitative continue (1)

- Référence: densité de probabilité



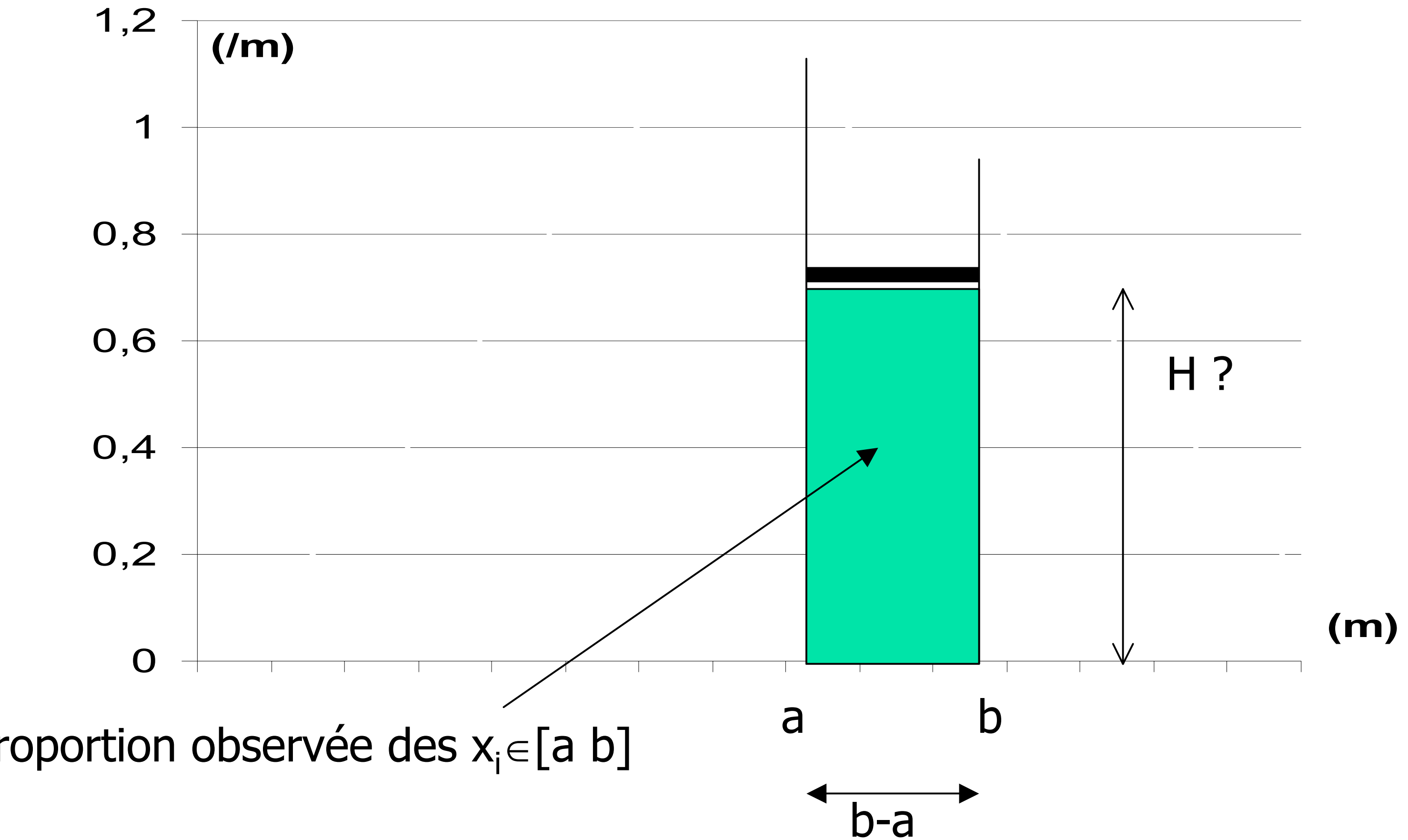
Représentation des résultats variable quantitative continue (2)

- Référence: densité de probabilité



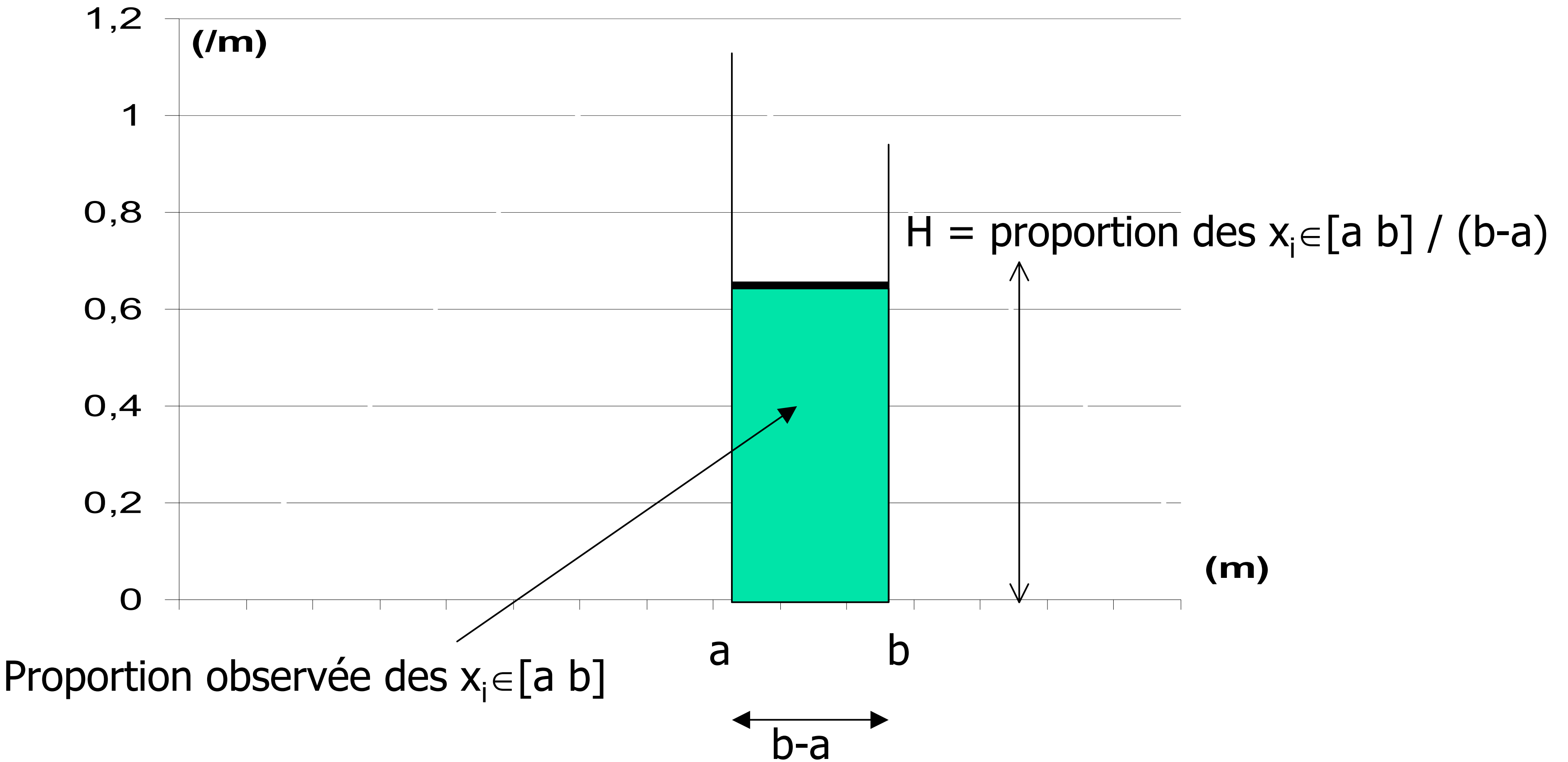
Représentation des résultats variable quantitative continue (3)

- Référence: densité de probabilité



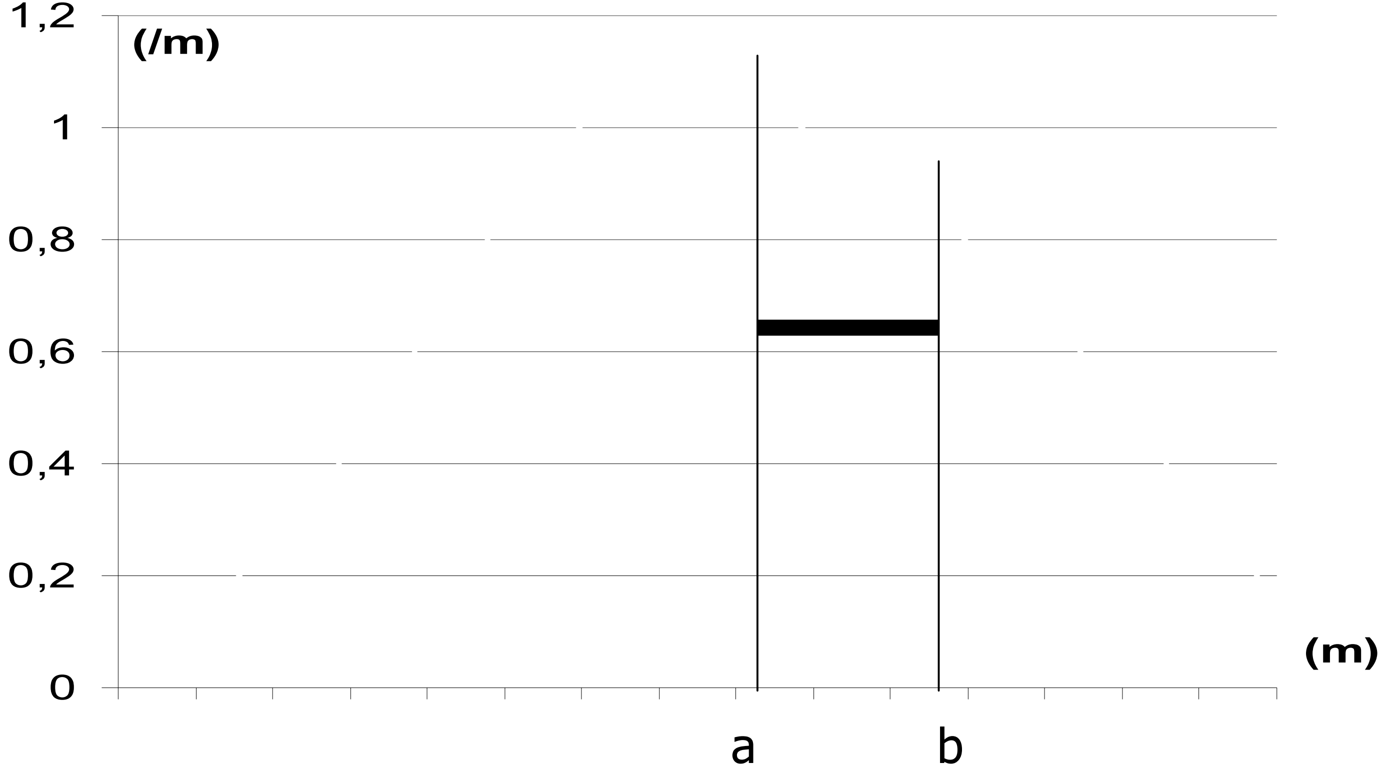
Représentation des résultats variable quantitative continue (4)

- Référence: densité de probabilité



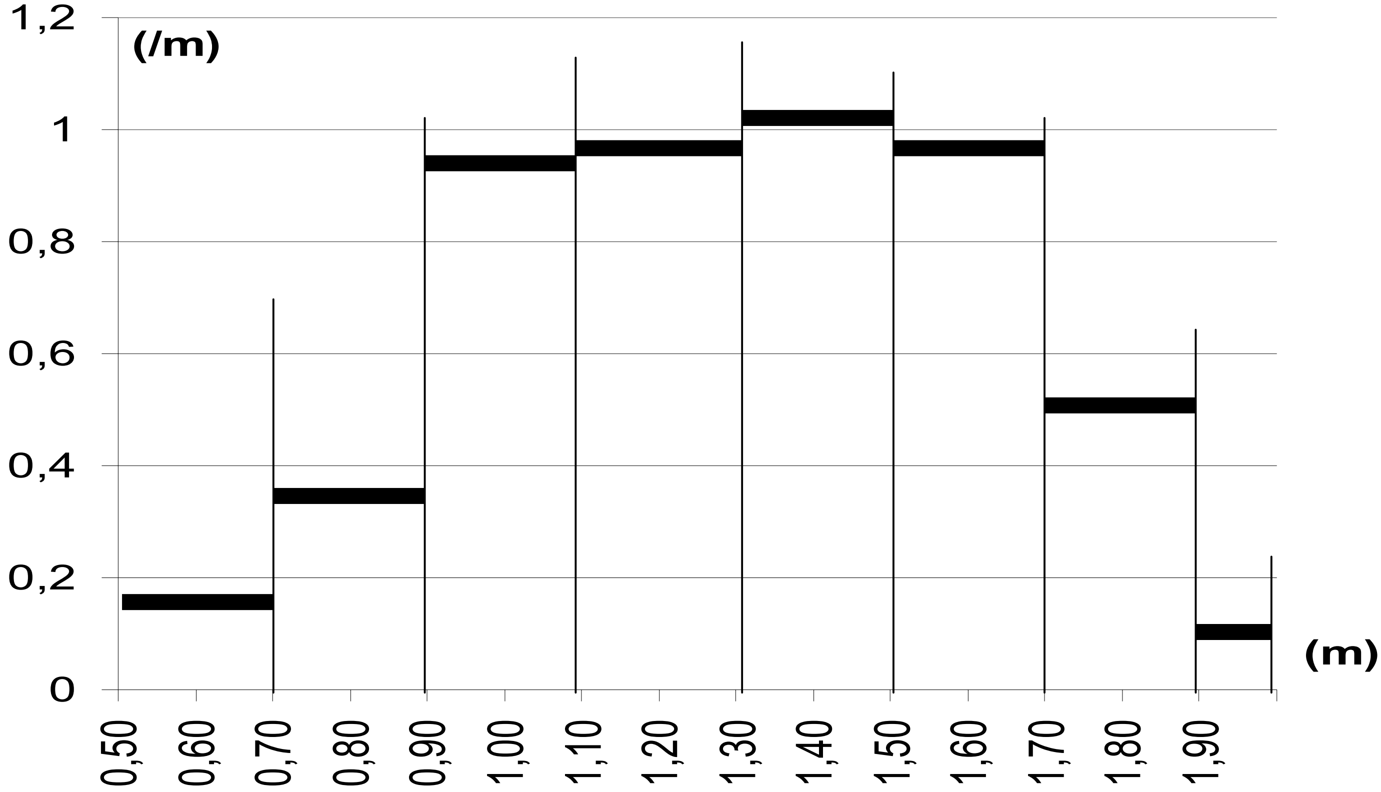
Représentation des résultats variable quantitative continue (5)

- Référence: densité de probabilité



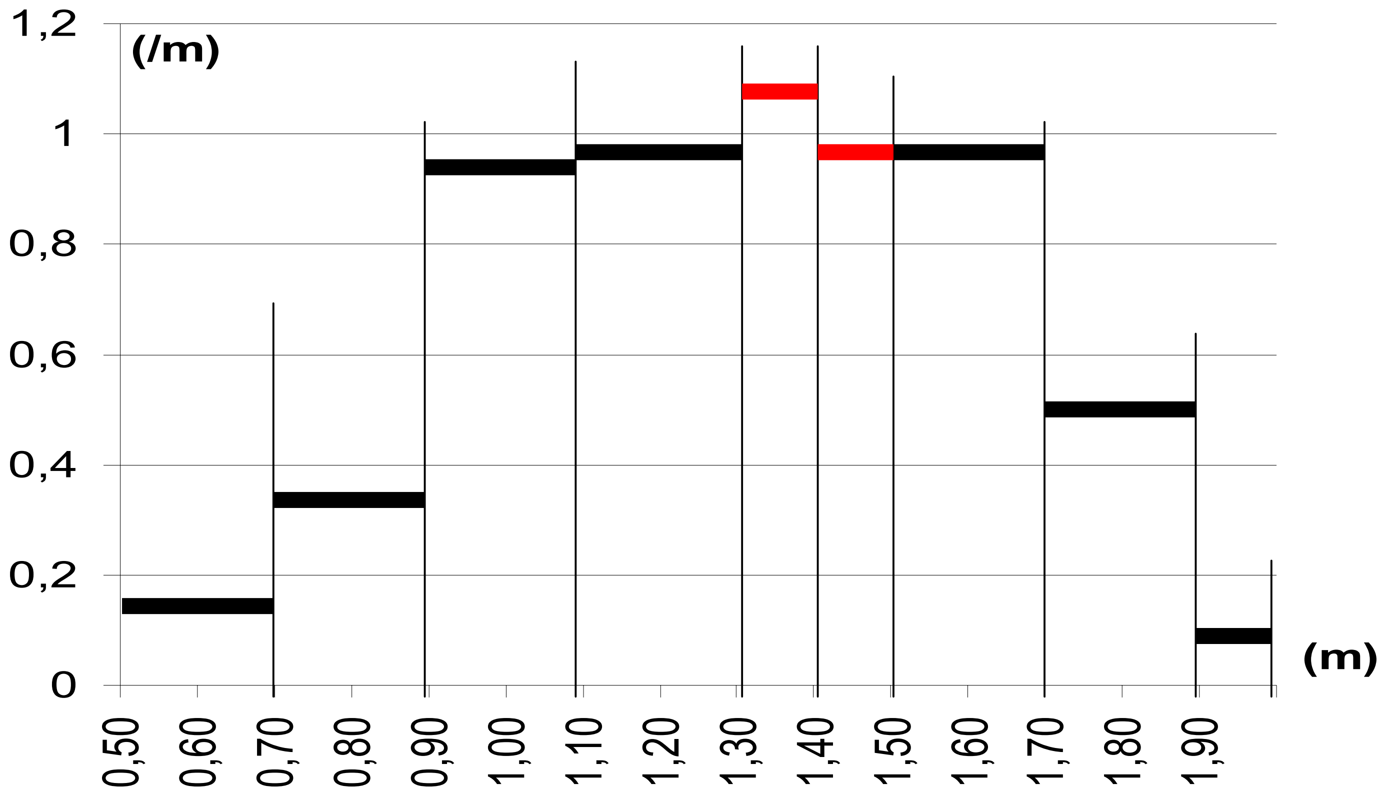
Représentation des résultats variable quantitative continue (6)

- Référence: densité de probabilité



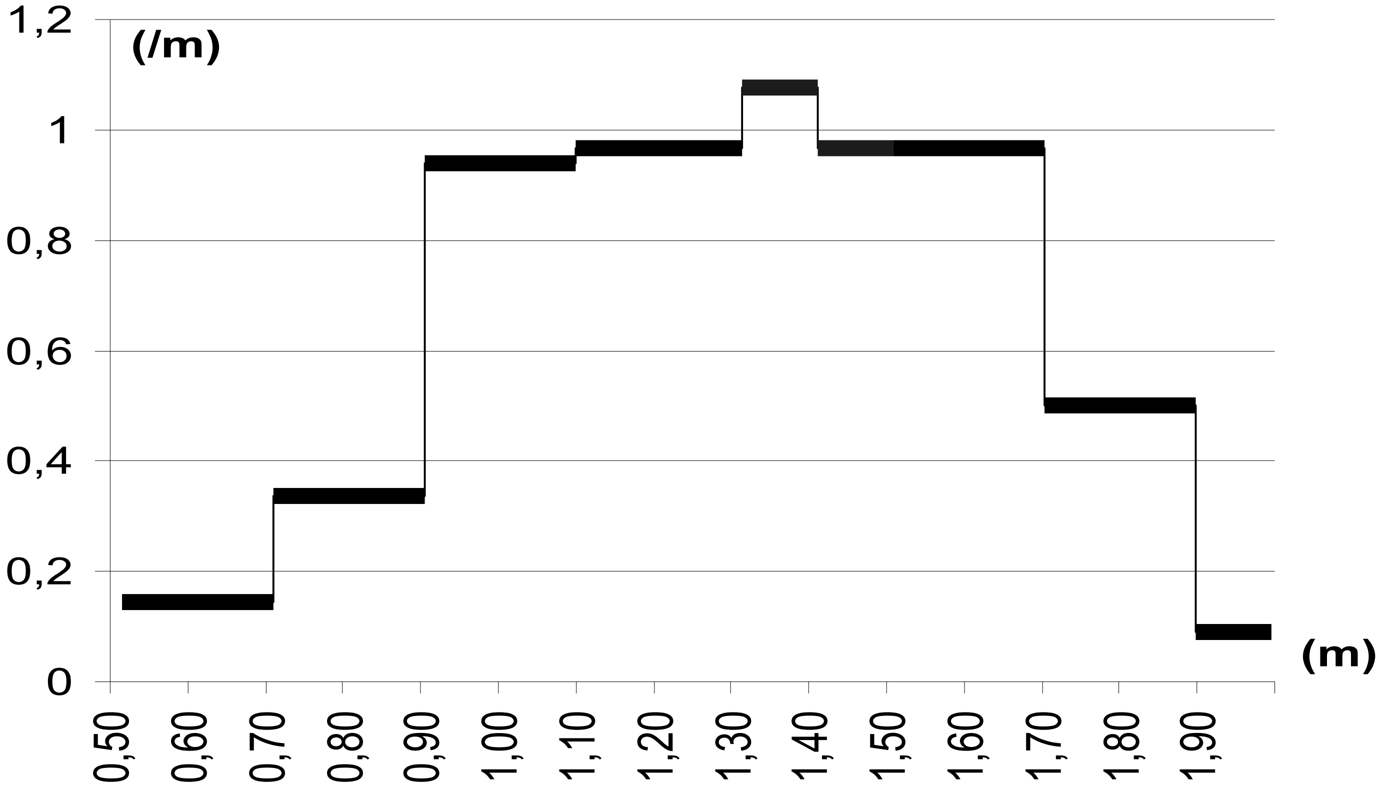
Représentation des résultats variable quantitative continue (7)

- Référence: densité de probabilité



Représentation des résultats variable quantitative continue (8)

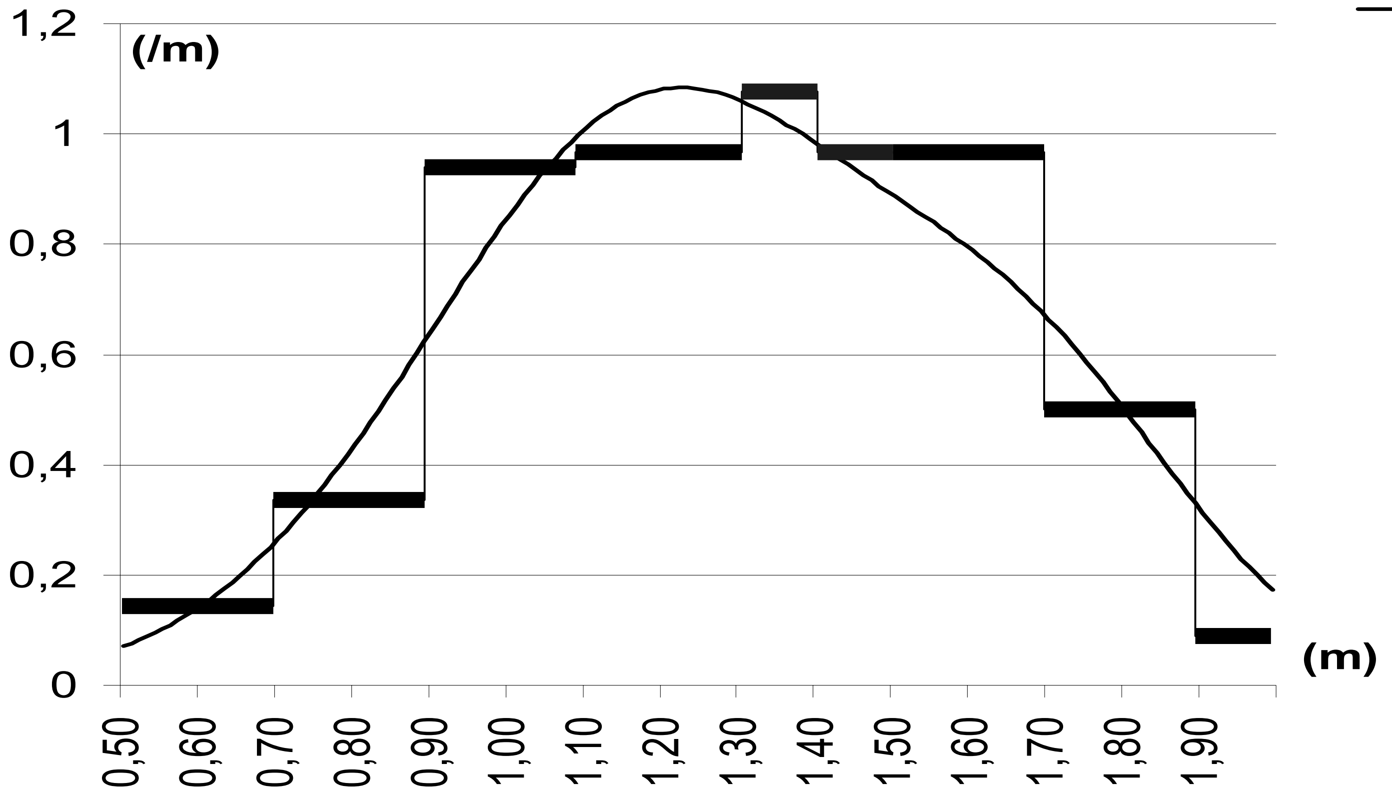
- Référence: densité de probabilité



•Réponse : HISTOGRAMME

Représentation des résultats variable quantitative continue (9)

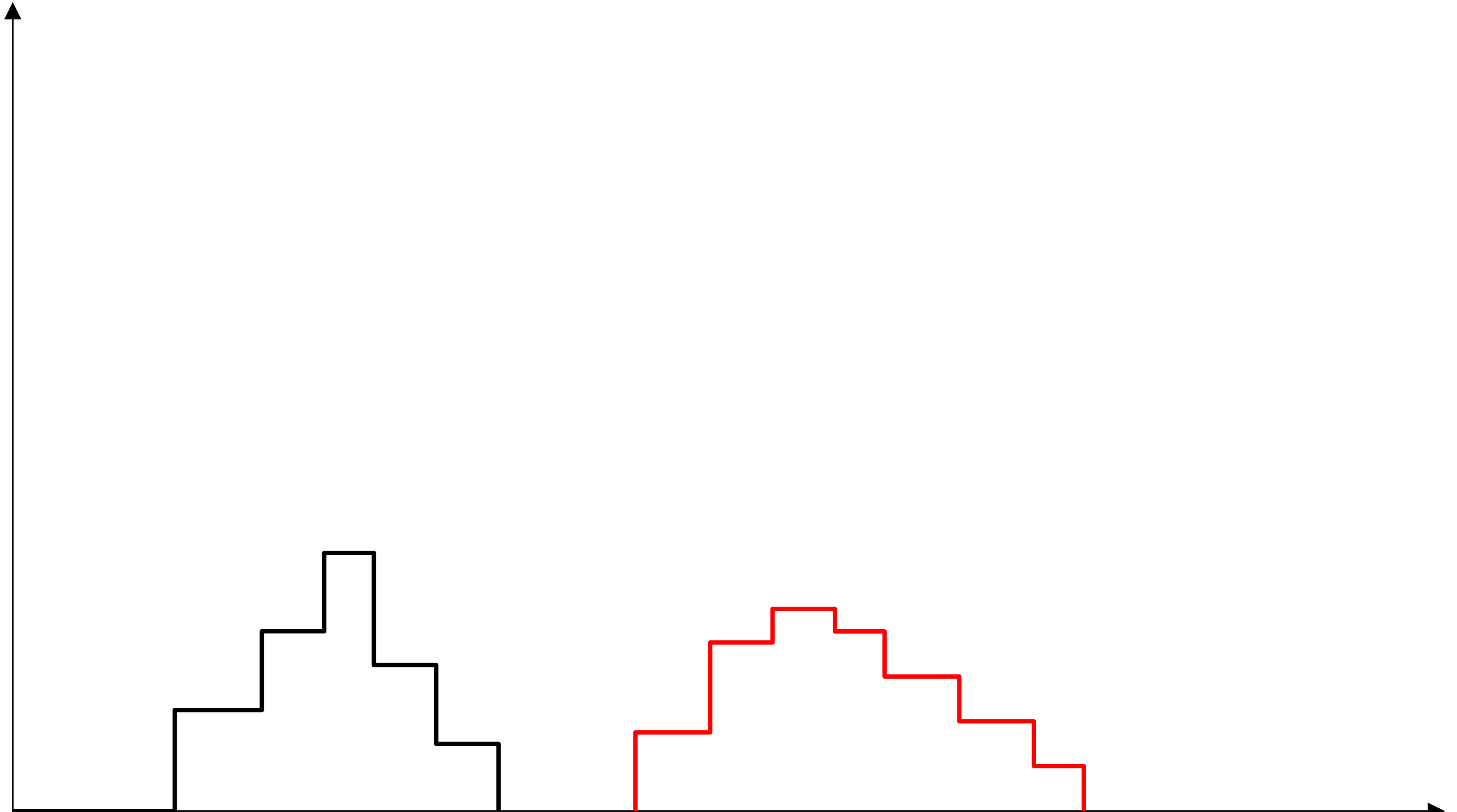
- Référence: densité de probabilité

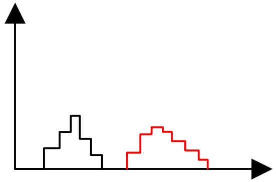


•Réponse : HISTOGRAMME

Représentation simplifiée des résultats variables quantitatives

- Localisation des valeurs
  - Médiane observée: valeur telle qu'il y ait autant de valeurs qui lui sont supérieures que de valeurs qui lui sont inférieures
  - Moyenne observée: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Remarque: m est une réalisation de la variable aléatoire $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ appelée variable aléatoire moyenne arithmétique construite sur un échantillon de taille n de X.

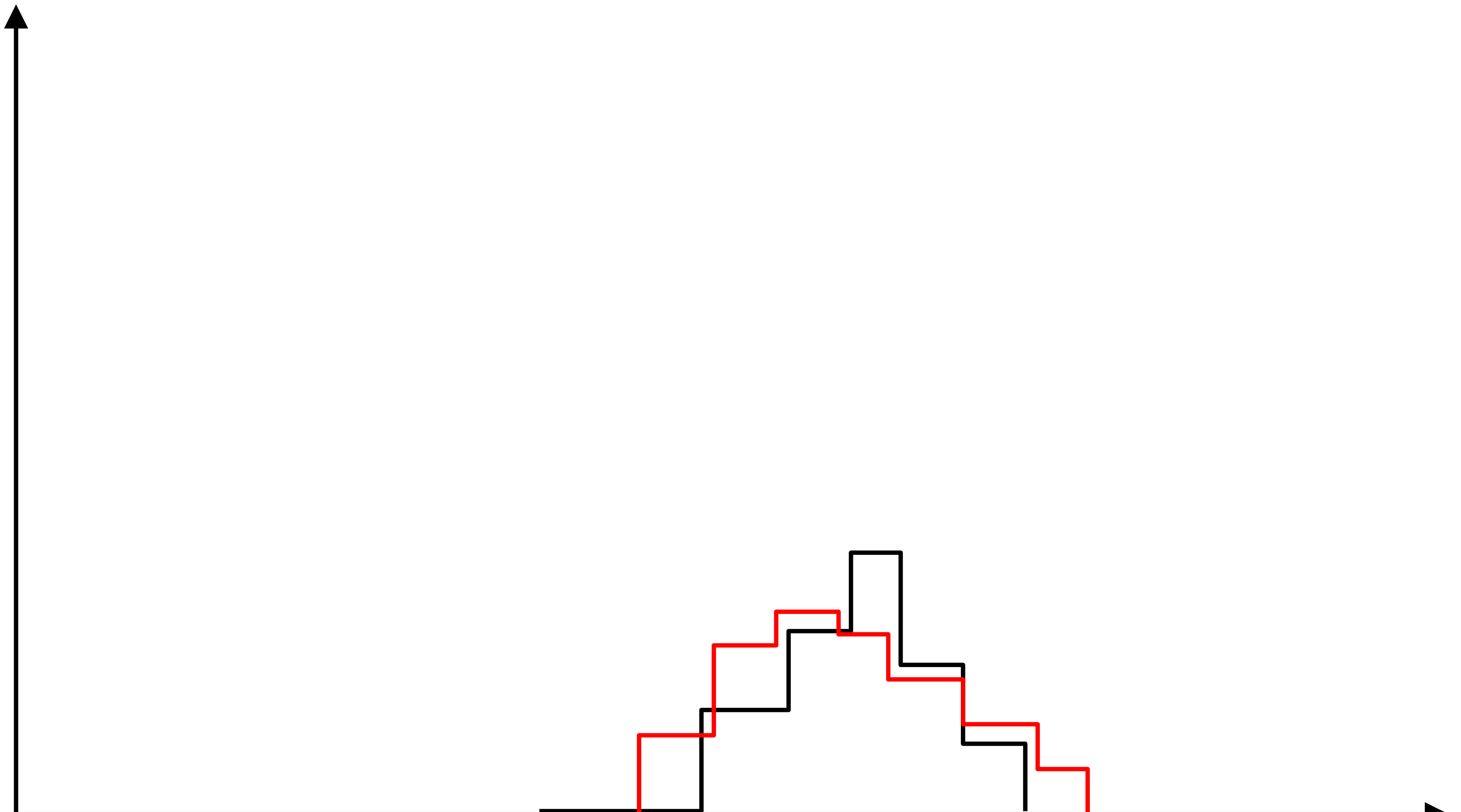


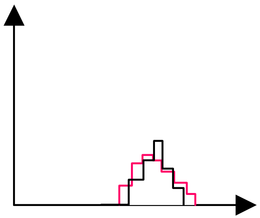


Représentation simplifiée des résultats variables quantitatives

- Dispersion des valeurs:
  - variance observée $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$
  - Écart type observé: $\sqrt{s^2}$ noté s
  - Remarque: s² est une réalisation de la variable aléatoire variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$





Reformulation de la moyenne observée (1)

- X variable aléatoire discrète finie à k valeurs possibles: val_1, val_2, …., val_k (ex: jeu de dé)
- Toute valeur de l'échantillon de valeurs, x_i , égale val_j pour un j bien choisi.
- Alors:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} \frac{n_j}{n} val_j$$

Reformulation de la moyenne observée (2)

- Exemple

$$m = \frac{1}{8}(2+4+6+3+3+1+4+5)$$

$x_1=val_2;....; x_3=val_6 ;......$

$n_1=1, n_2=1, n_3=2, n_4=2, n_5=1, n_6=1$

$$m = \frac{1}{8}(1*1+1*2+2*3+2*4+1*5+1*6)$$

Reformulation de la moyenne observée (3)

- Donc: $$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} \frac{n_j}{n} val_j$$
- Proche de : $$\sum_{j=1}^{k} val_j \Pr(X = val_j)$$
- Qui est l'espérance mathématique de X, E(X): μ ->synonymes
  - Espérance mathématique
  - Moyenne 'vraie'
  - Moyenne théorique
- $\boxed{\text{Ce résultat vaut pour toute variable aléatoire}}$

Reformulation de la variance observée

- De la même façon: la variance observée

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

approche

$$\sum_{j=1}^{k}(\mathrm{val}_j - \mu)^2 \ \Pr(X = \mathrm{val}_j)$$

- qui est la variance de X, $E(X-\mu)^2 : \sigma^2$
-> synonymes
  - variance
  - variance 'vraie'
  - variance théorique
- $\boxed{\text{Ce résultat vaut pour toute variable aléatoire}}$

Cas d'une variable de Bernoulli (1)

- Variable très utilisée en médecine
  - à la frontière qualitative-quantitative
  - codage 0-1 des modalités
  - E (X) = Pr (X=1) $=\pi =$ probabilité (un individu présente la modalité 1)
- Interprétation de la moyenne observée
  - m $= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(0+1+0+0+1+1+...) =$

$$\frac{\text{nombre de fois où } X=1}{n} = \text{proportion observée}$$

Cas d’une variable de Bernoulli (2)

- la proportion observée, notée p, est une moyenne observée
- la proportion observée p approche E(X) donc approche π
- π pourra s’appeler proportion ‘vraie’

Cas d’une variable de Bernoulli (3)

- Rappel : m est une réalisation de M_n
- Si X est une variable de Bernoulli, M_n sera noté P_n et on parlera de variable aléatoire ~~moyenne arithmétique~~ proportion construite sur un échantillon de taille n
- Avec ces notations adaptées, p est une réalisation de P_n.

Conclusion

- Ces premiers résultats (m approche E(X); p approche π) engagent à étudier les variables M_n et P_n.
- Objet de la suite.

La variable aléatoire moyenne arithmétique
Cours 6

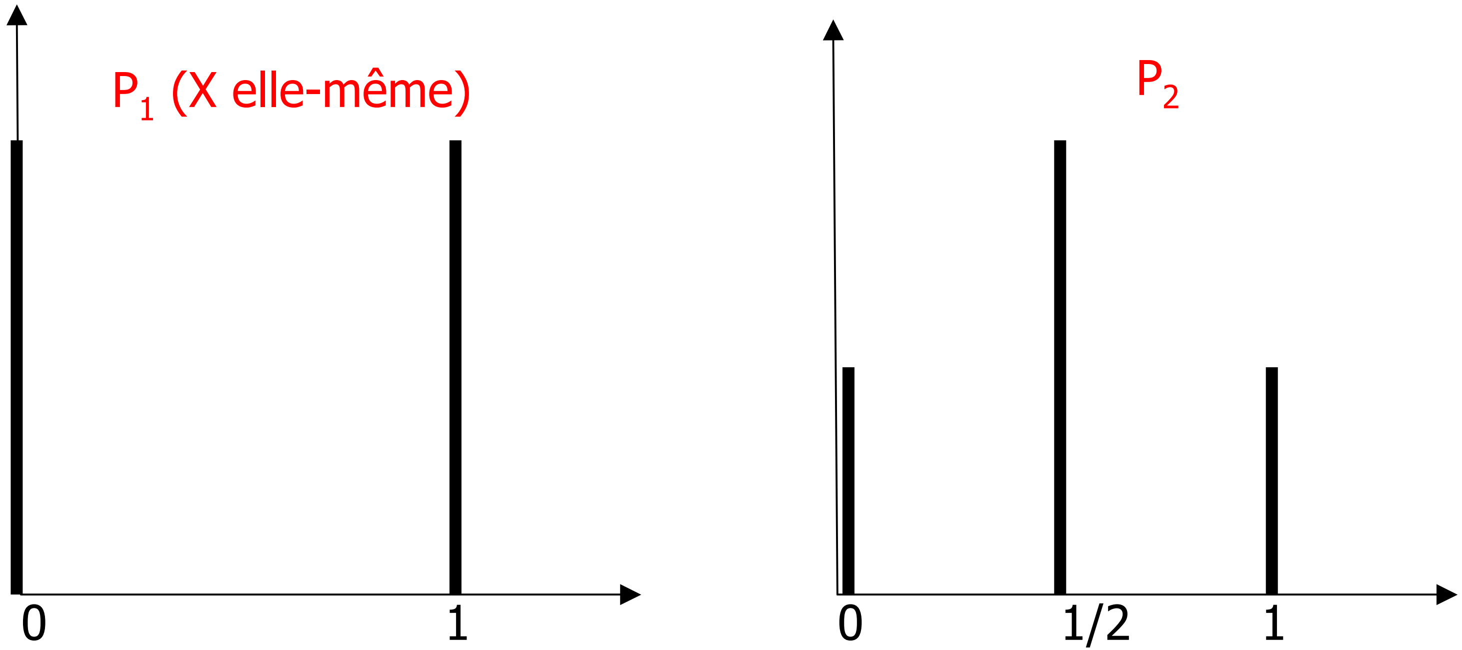
Etude de la variable aléatoire moyenne arithmétique (M_n et P_n)

- Variable étudiée:X
- Remarque: $M_1=P_1=X$
- Exemple: la variable P_2 ; elle repose sur:
  - une variable de Bernoulli: X. Ici on choisit $\pi=1/2$
  - un échantillon de taille 2: X_1,X_2

X_1	X_2	P_2	Probabilité	
0	0	½(0+0) = 0	1/4	[$(1-\pi)^2$ en général]
0	1	½(0+1) =1/2	1/4	
1	0	½(1+0) =1/2	1/4	
1	1	½(1+1) =1	1/4	

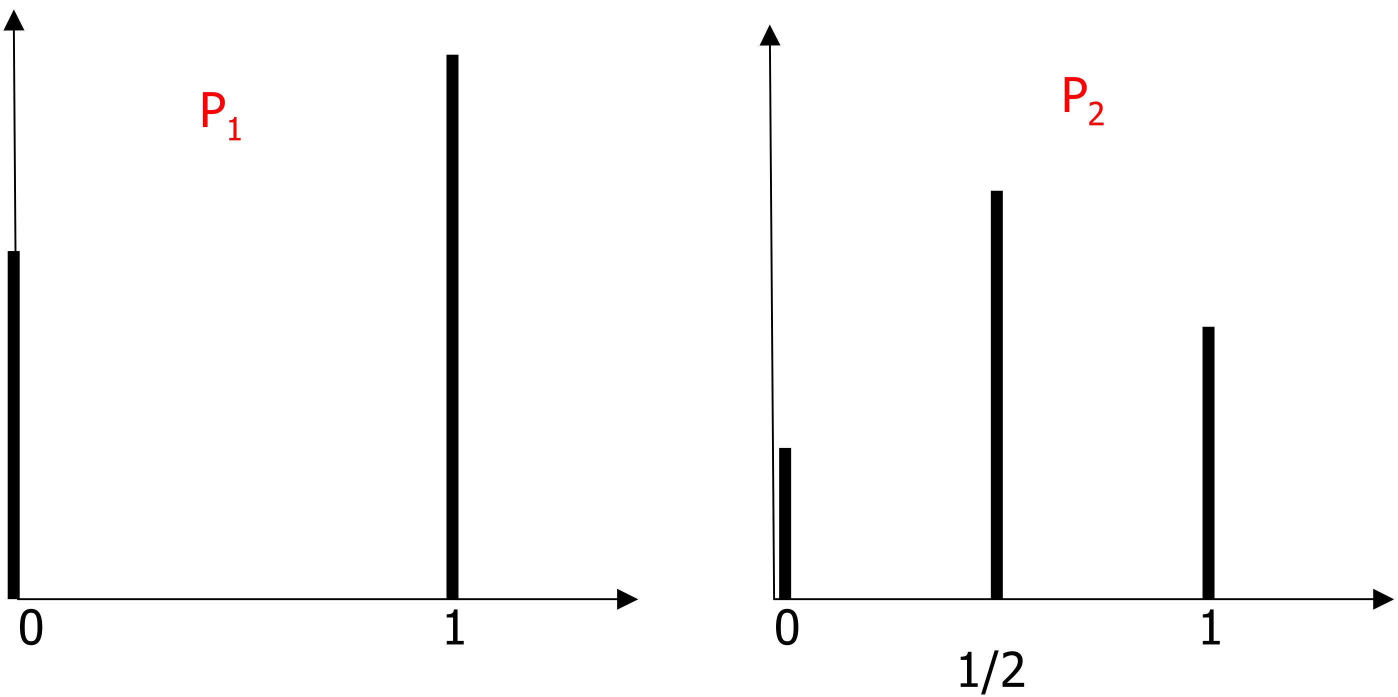
La variable aléatoire P_2 (suite)

- Répartition si $\pi=1/2$



La variable aléatoire P_2 (suite)

- Répartition si $\pi=0,4$



La variable aléatoire P_2 (suite)

En particulier:

- $E(P_2) = val_1*Pr(P_2=val_1)+val_2*Pr(P_2=val_2)+val_3*Pr(P_2=val_3)$

  $= 0 * 1/4 \quad + 1/2*1/2 \quad + 1*1/4$

  $=1/2=E(X)$ Généralement: $=\pi$

- $var(P_2)=E[(P_2-1/2)^2]=(0-1/2)^2*1/4+ (1/2-1/2)^2*1/2 +(1-1/2)^2 *1=1/8$

  $=1/2*1/4=1/2\ var(X)$ Généralement: $= 1/2\ \pi(1-\pi)$

Propriété de M_n (donc de P_n) (2)

- Résultat général:

$$\boxed{\begin{aligned} &E(M_n) = E(X) \\ &var(M_n) = \frac{1}{n} var(X) \qquad ; \text{donc } var(\sqrt{n}\, M_n) = var\,(X) \end{aligned}}$$

- Attendu en fait car:

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = E(X)$$

$$var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X) = \frac{1}{n} var(X)$$

Propriété de M_n (donc de P_n) (1)

- Habillage dans le cas d'une proportion P_n:

$E(P_n) = \pi$

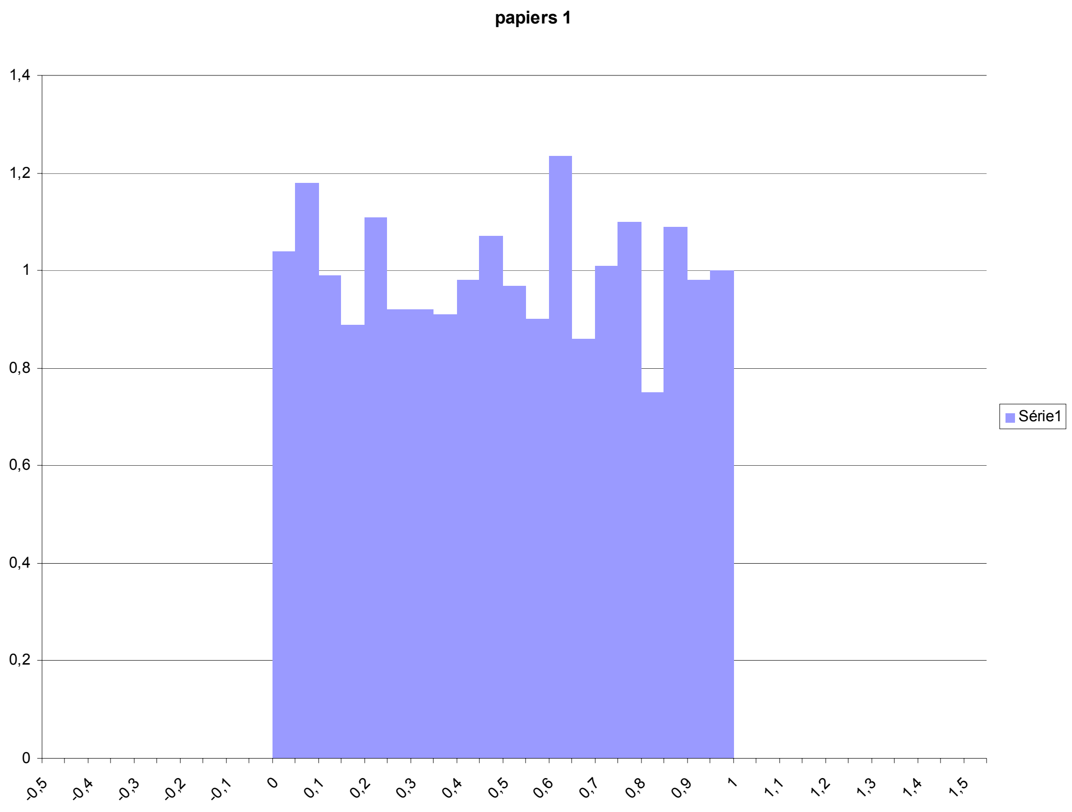
var(X)

$$\mathrm{var}(P_n) = \frac{1}{n}\boxed{\pi(1-\pi)}$$

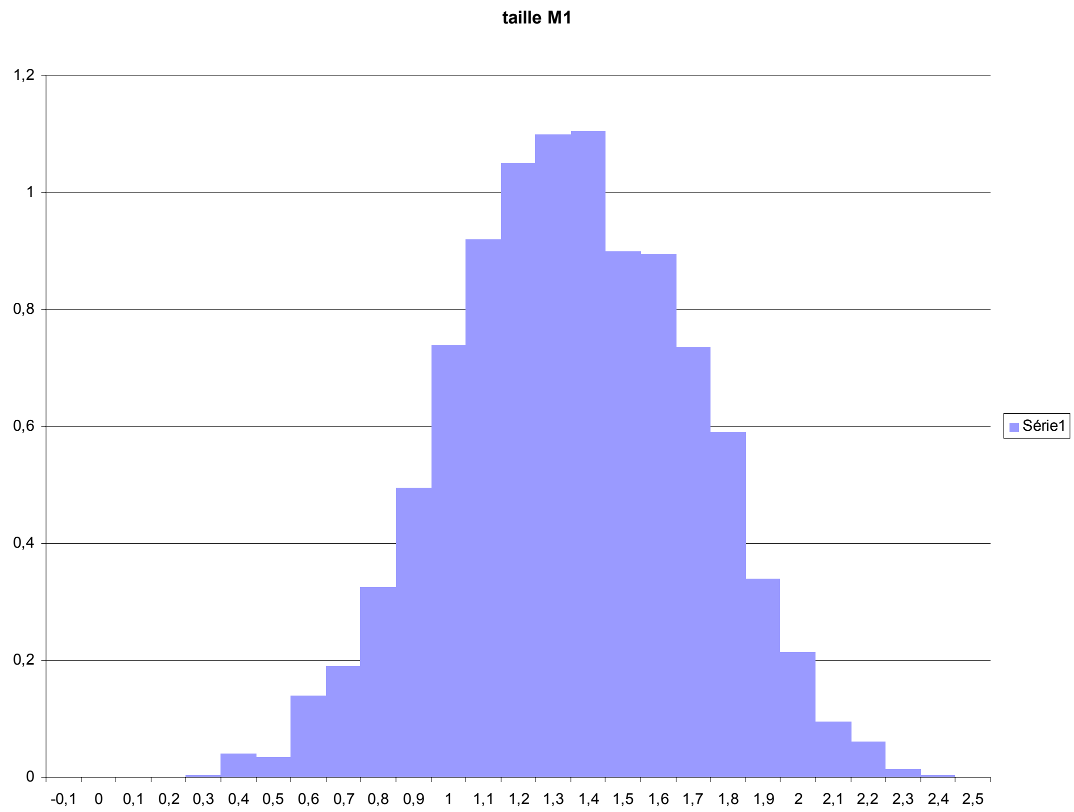
Une étude ambitieuse

- Etude simultanée des variables moyennes arithmétiques, M_n et Q_n construites sur 2 variables aléatoires X et Y
  - X: nombres entre 0 et 1 écrits sur des papiers contenus dans une urne
  - Y: taille des adultes dans la population humaine
- n=1: étude pratique de M_1 (X) et Q_1 (Y)
  - Constitution d'un échantillon de chaque variable
  - Construction des histogrammes

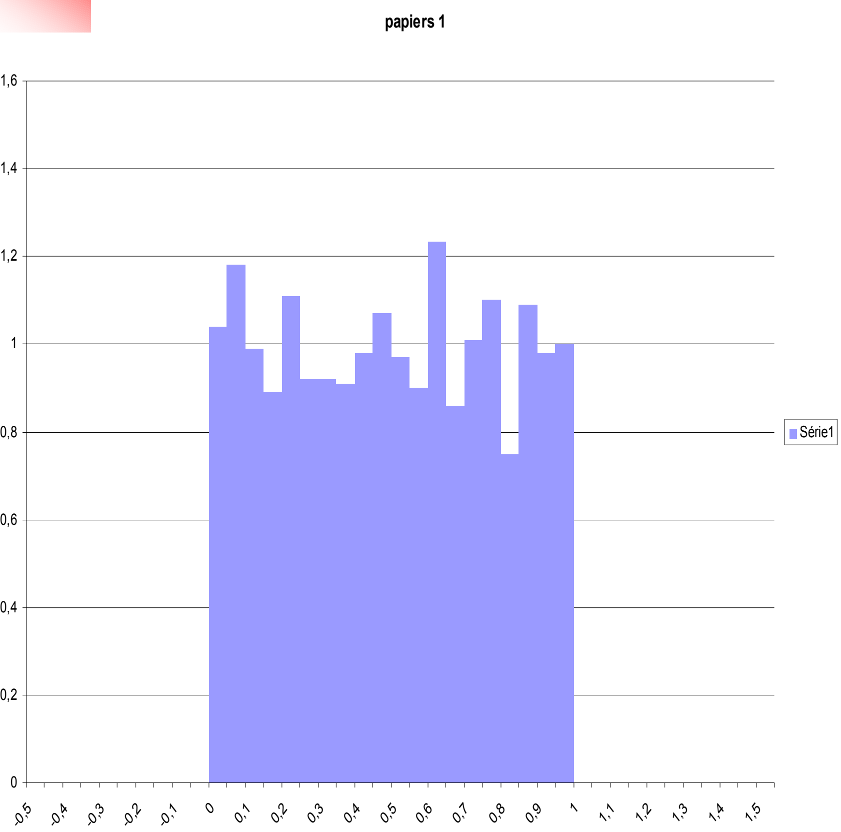
Histogramme de M_1 (X)

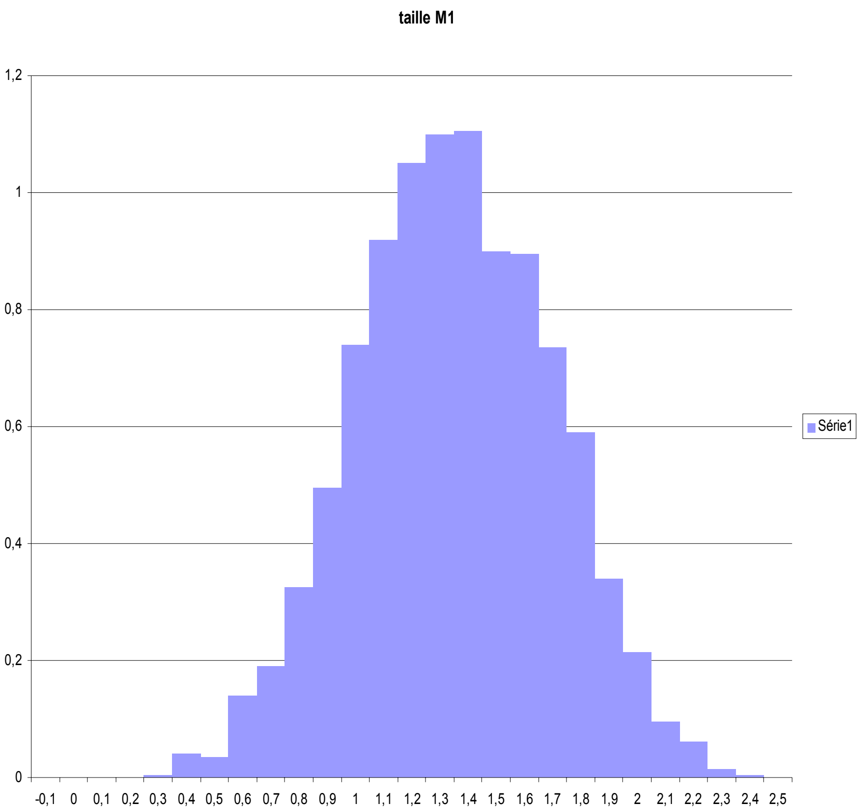


Histogramme de Q1 (Y)

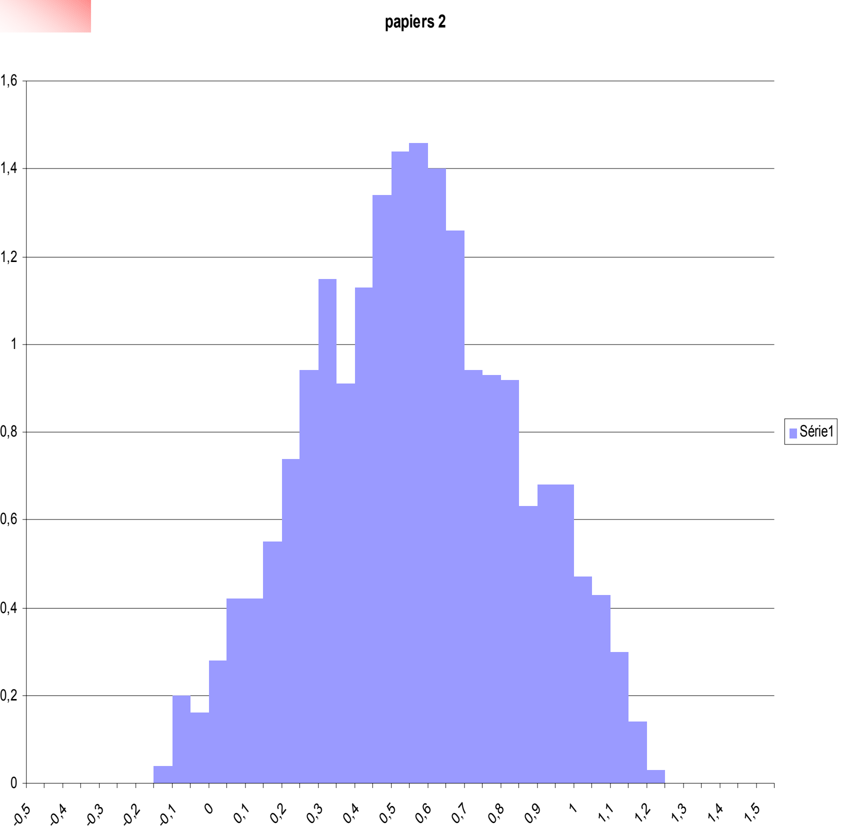


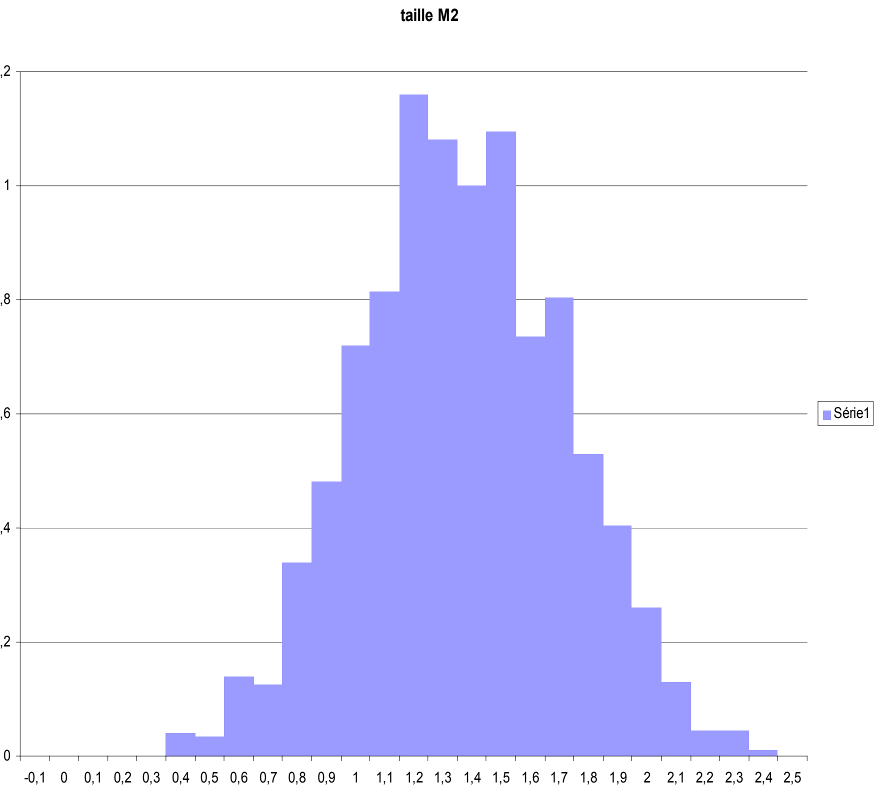
n=1: histogrammes de M1 (X) et Q1 (Y)





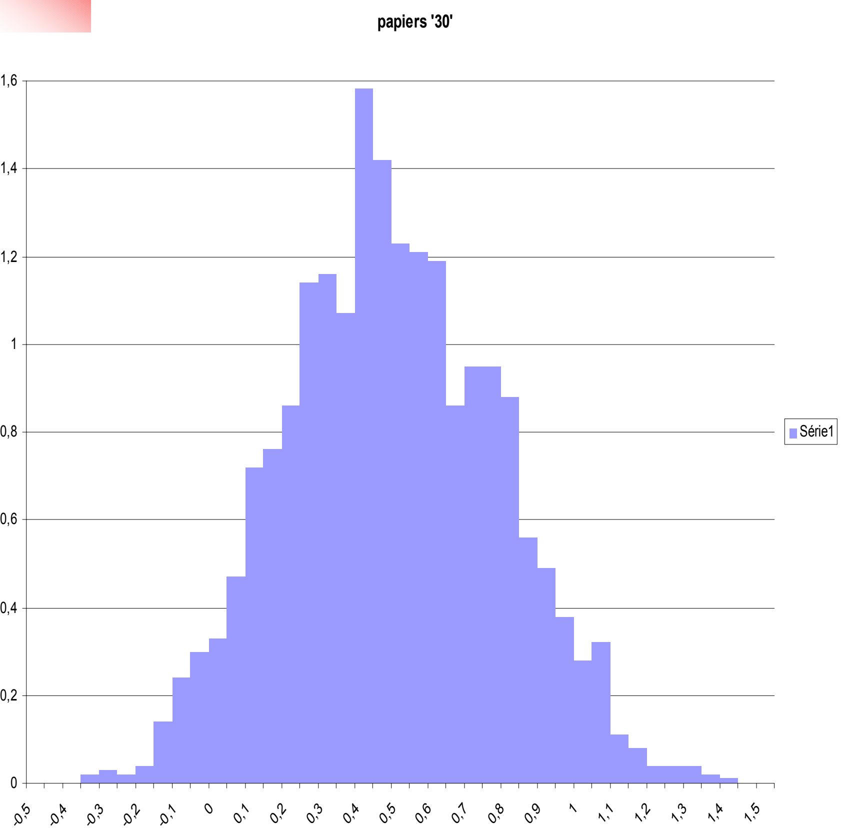
n=2: histogrammes de M_2 et Q_2 (*)

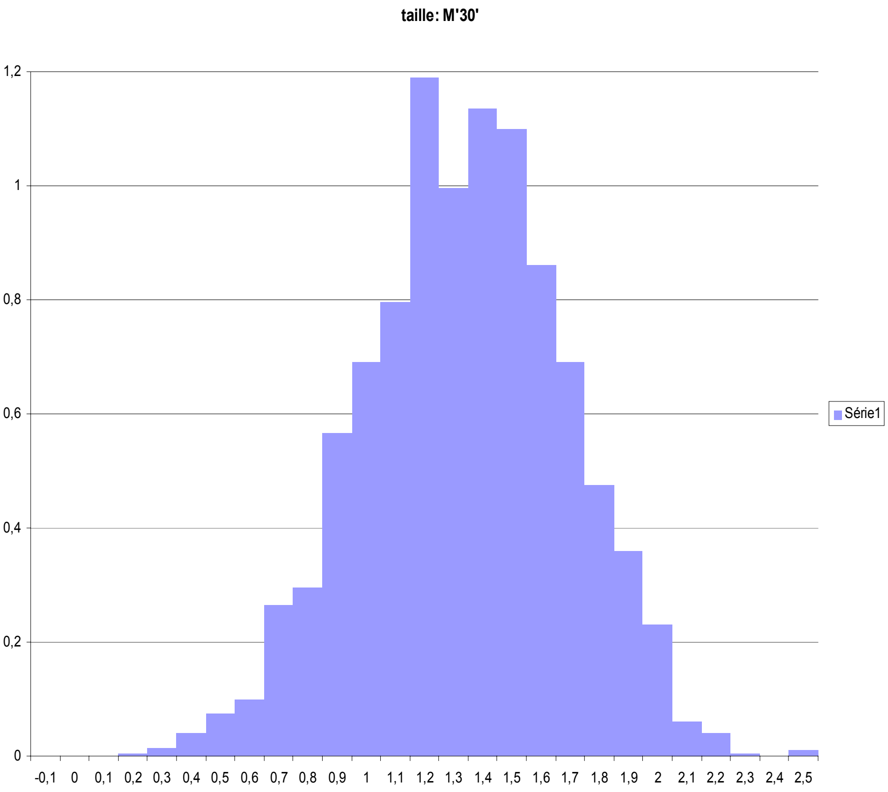




* en réalité de $\sqrt{2}M_2$ et $\sqrt{2}Q_2$, ramenés à moyennes observées 0,5 et 1,3

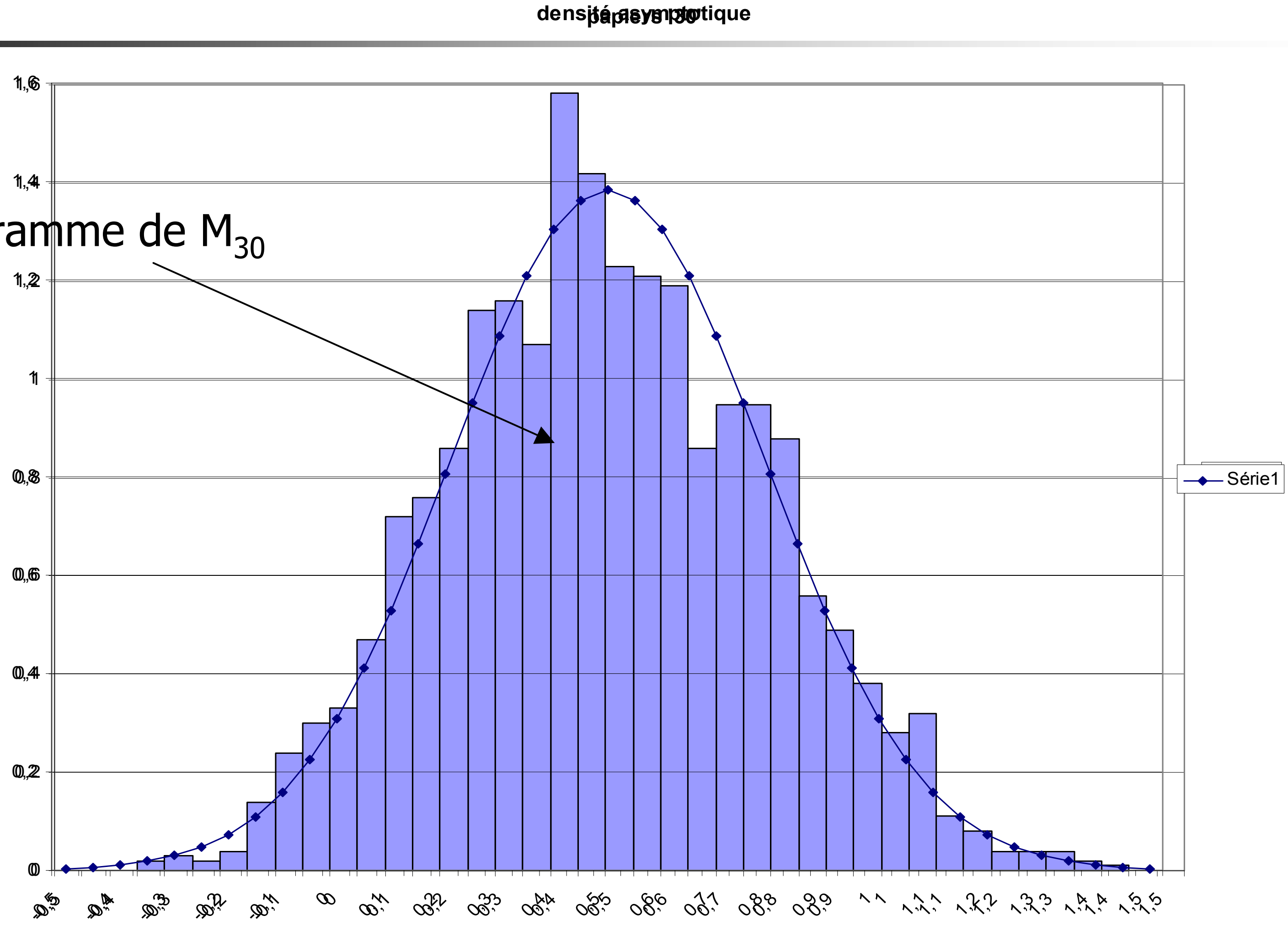
n=30: histogrammes de M_{30} et Q_{30} (*)





* en réalité de $\sqrt{30}M_2$ et $\sqrt{30}Q_2$, ramenés à moyennes observées 0,5 et 1,3

Histogramme de M_{30}



Le théorème central limite*** (Laplace,Gauss)

- Pour toute variable quantitative, la distribution limite (lorsque n croit indéfiniment) de:

$$\frac{M_n - E(X)}{\sqrt{\frac{var(X)}{n}}}$$

$E(M_n)$

$var(M_n)$

est la distribution d'une variable normale centrée réduite (N(0,1))

Le théorème central limite, version pratique (1)

- Lorsque n est ‘assez grand’, la variable

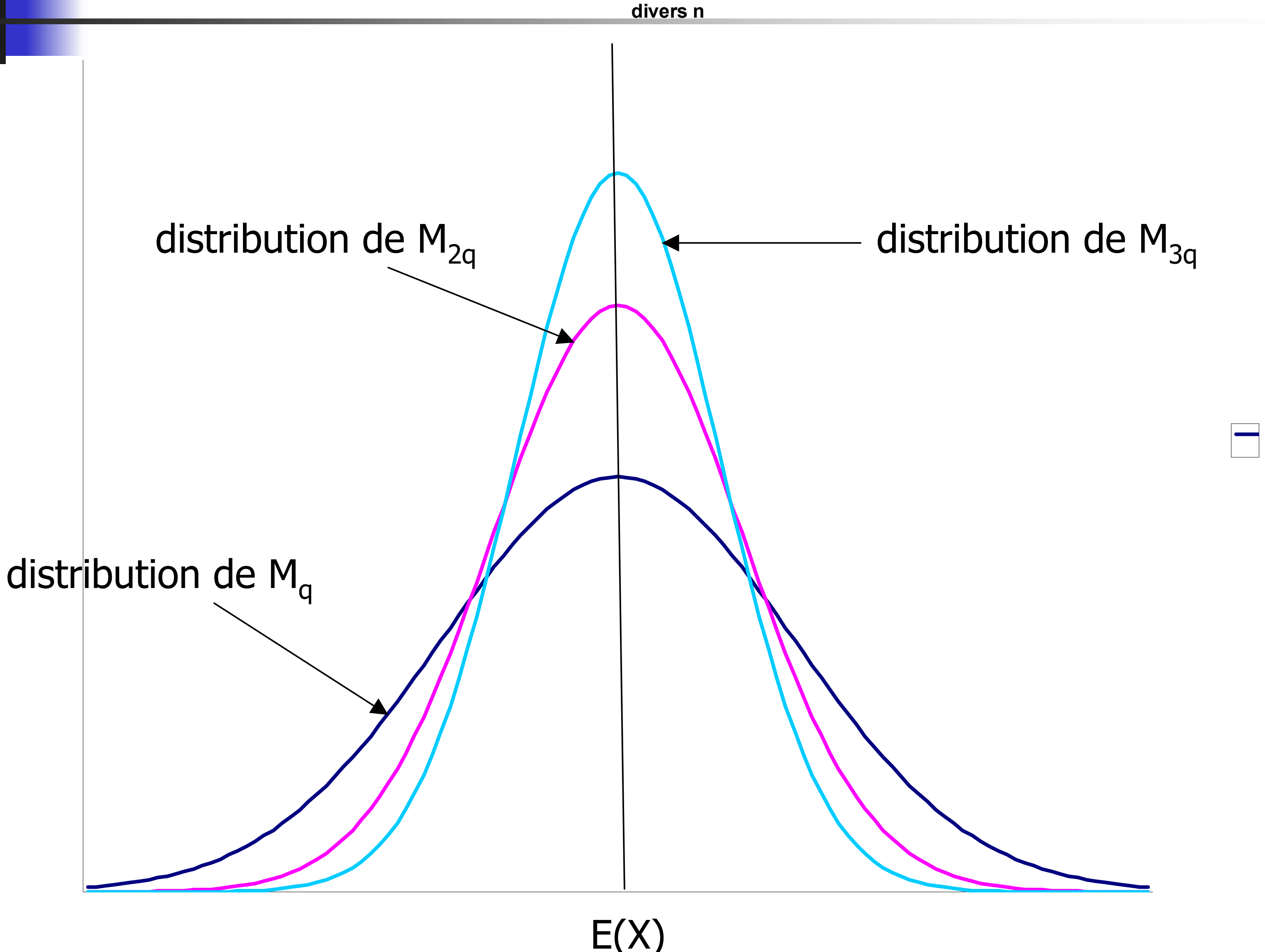
$$\frac{M_n - E(X)}{\sqrt{\frac{var(X)}{n}}}$$

  - a ‘à peu près’ une distribution d’une variable normale centrée réduite.
  - est ‘à peu près’ distribuée selon une loi normale centrée réduite

Le théorème central limite, version pratique (2)

- Lorsque n est 'assez grand'
  - $$\frac{M_n - E(X)}{\sqrt{\frac{\mathrm{var}(X)}{n}}} \underset{\text{à peu près}}{\sim} N(0,1)$$
  - $$M_n \underset{\text{à peu près}}{\sim} N(E(X), \frac{\mathrm{var}(X)}{n})$$

Diverses distributions de M_n pour divers n 'assez grands'



Le théorème central limite, version pratique (3)

- Qu'est-ce qu'assez grand ? Trois cas.

- 1. X est une variable normale.
Alors le théorème est vrai pour tout n
- 2. X n'est ni normale ni de Bernoulli.
Alors on considère que l'approximation est valide si $n \geq 30$
- 3. X est une variable de Bernoulli.
Alors l'approximation est valide si $n\pi \geq 5$ et $n(1-\pi) \geq 5$

Le théorème central limite, version pratique (4)

- Forme habillée du théorème dans ce dernier cas:

  Lorsque $n\pi \geq 5$ et $n(1-\pi) \geq 5$, la variable

$$\frac{P_n - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \underset{\text{à peu près}}{\sim} N(0,1)$$

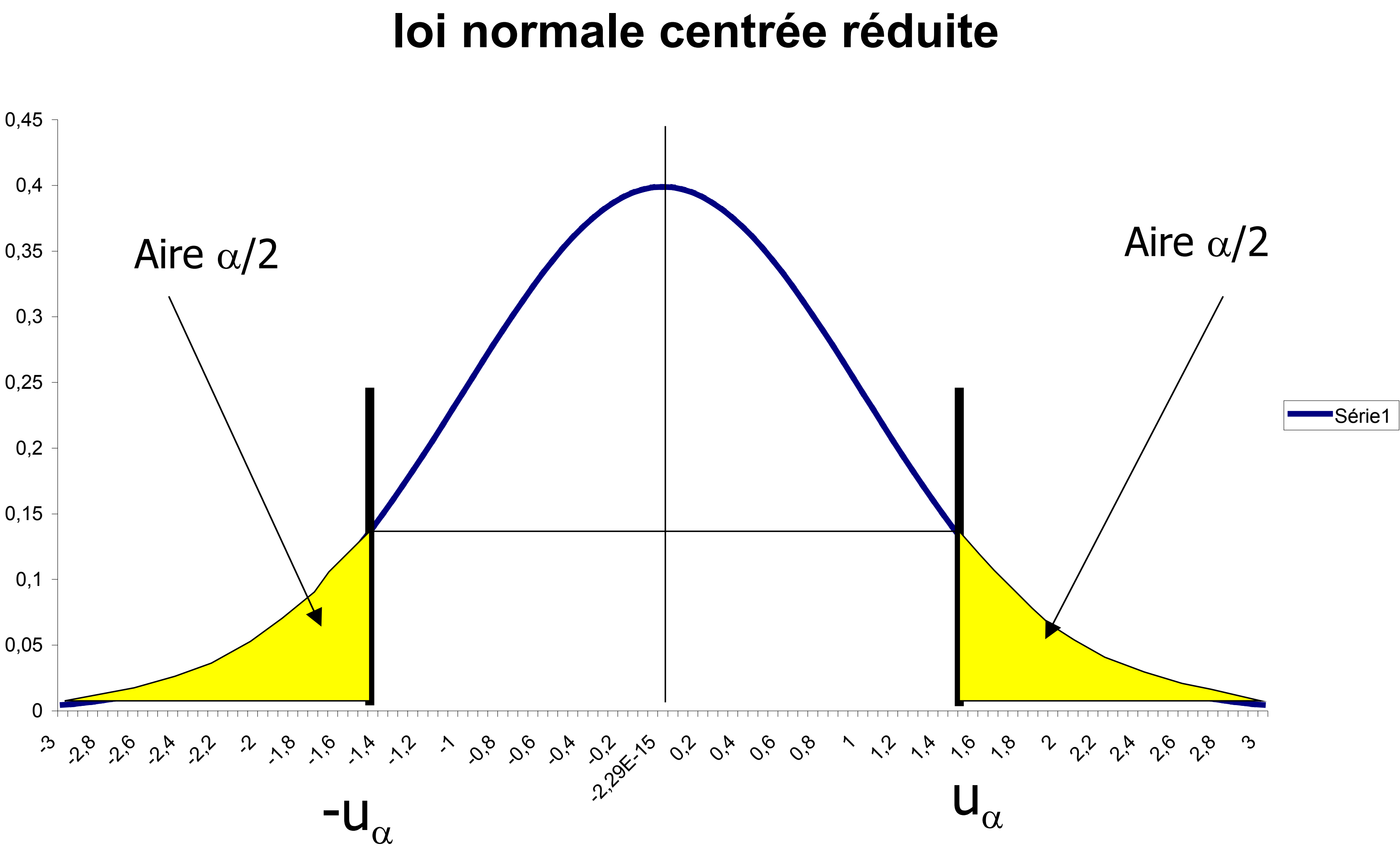
Le rôle majeur de N(0,1) (1)

- vaut bien une convention: la convention u_α
  - Soit X distribuée N(0,1)
  - Pour toute valeur α entre 0 et 1, u_α est la valeur telle que

$$\Pr(X \notin [-u_\alpha \; u_\alpha]) = \alpha$$

  - Le calcul de u_α connaissant α, ou l'inverse, se fait à l'aide d'une table numérique

Le rôle majeur de N(0,1) (2)



Application du théorème central limite (1)

- Problème: déterminer [a b] tel que

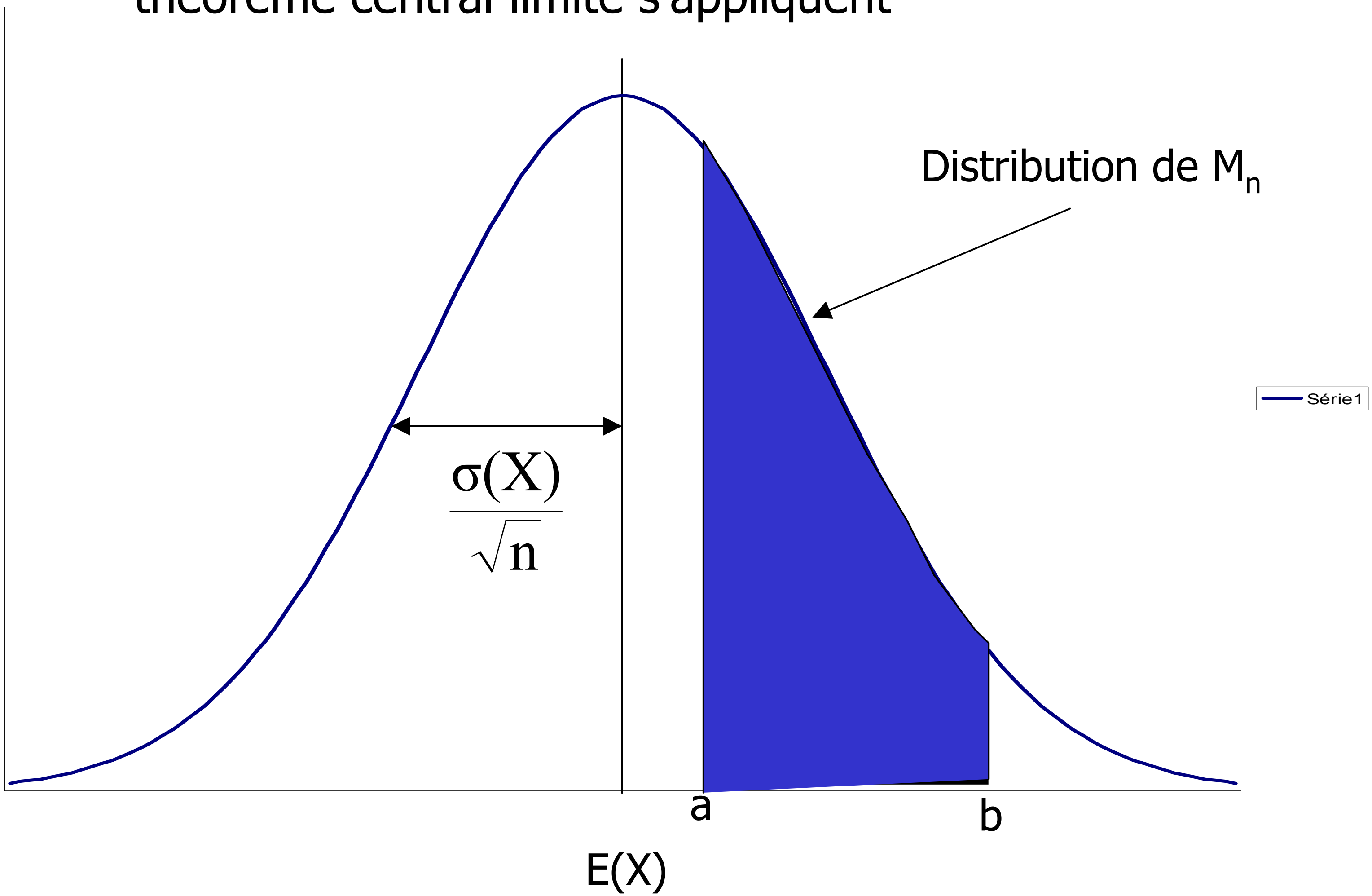
$$\Pr(M_n \in [a\ b]) = 1 - \alpha$$

pour α choisi (exemple 0,3)

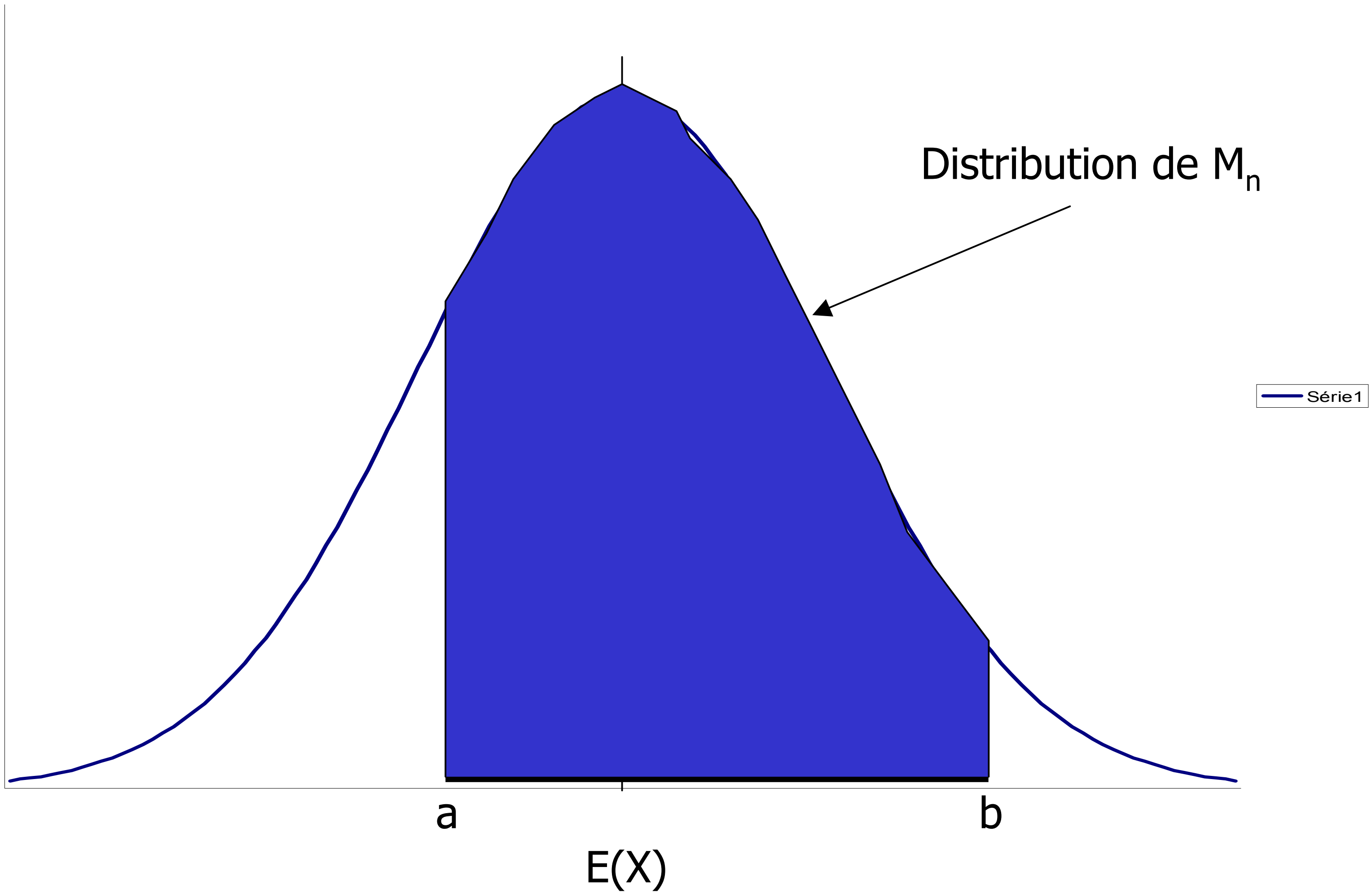
- Un tel intervalle s'appelle INTERVALLE DE PARI pour la variable aléatoire M_n
  - de niveau 1-α
  - au risque α

Application du théorème central limite (2)

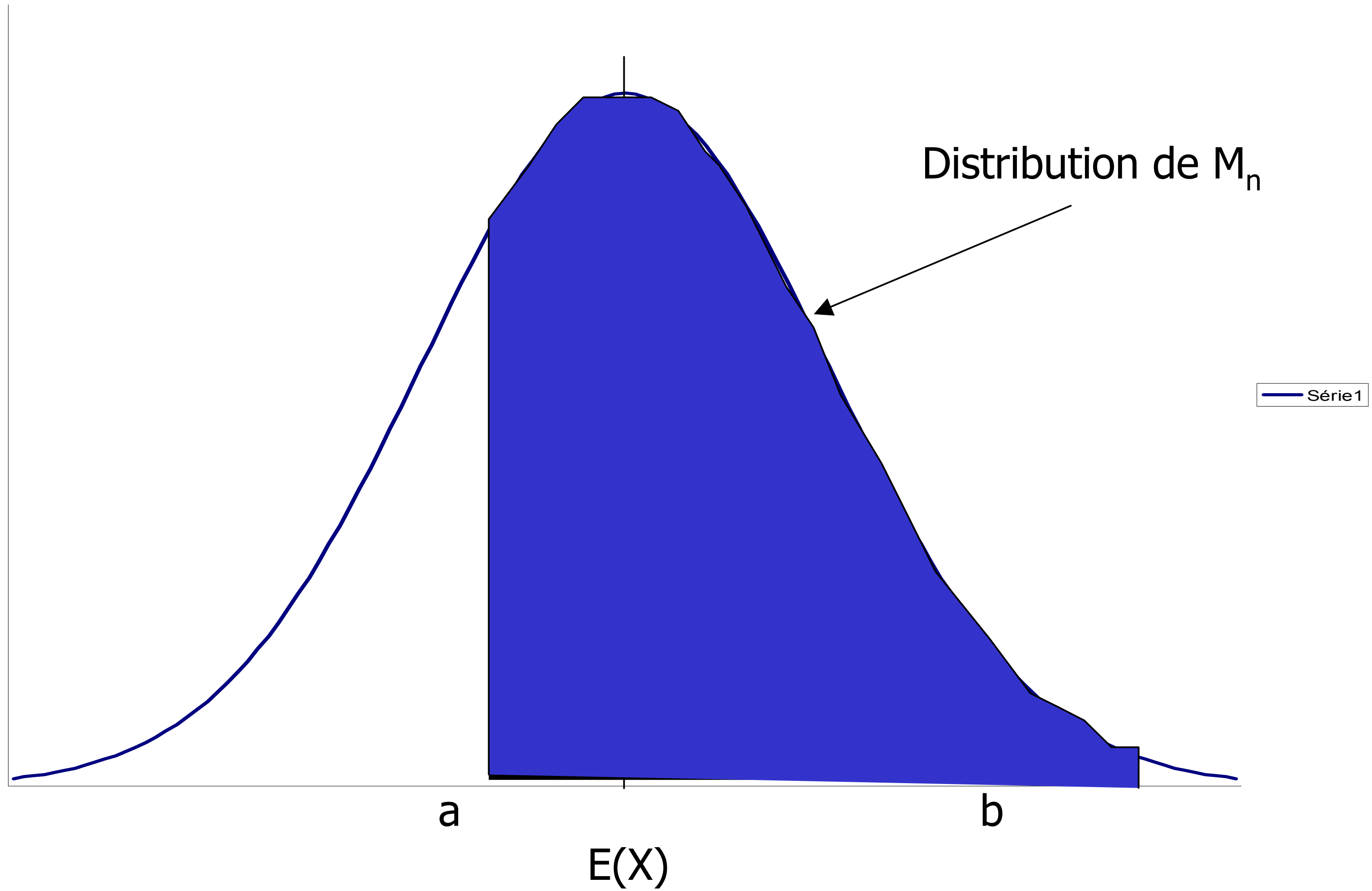
- Solution: on suppose que les conditions d'application du théorème central limite s'appliquent



Application du théorème central limite (3)

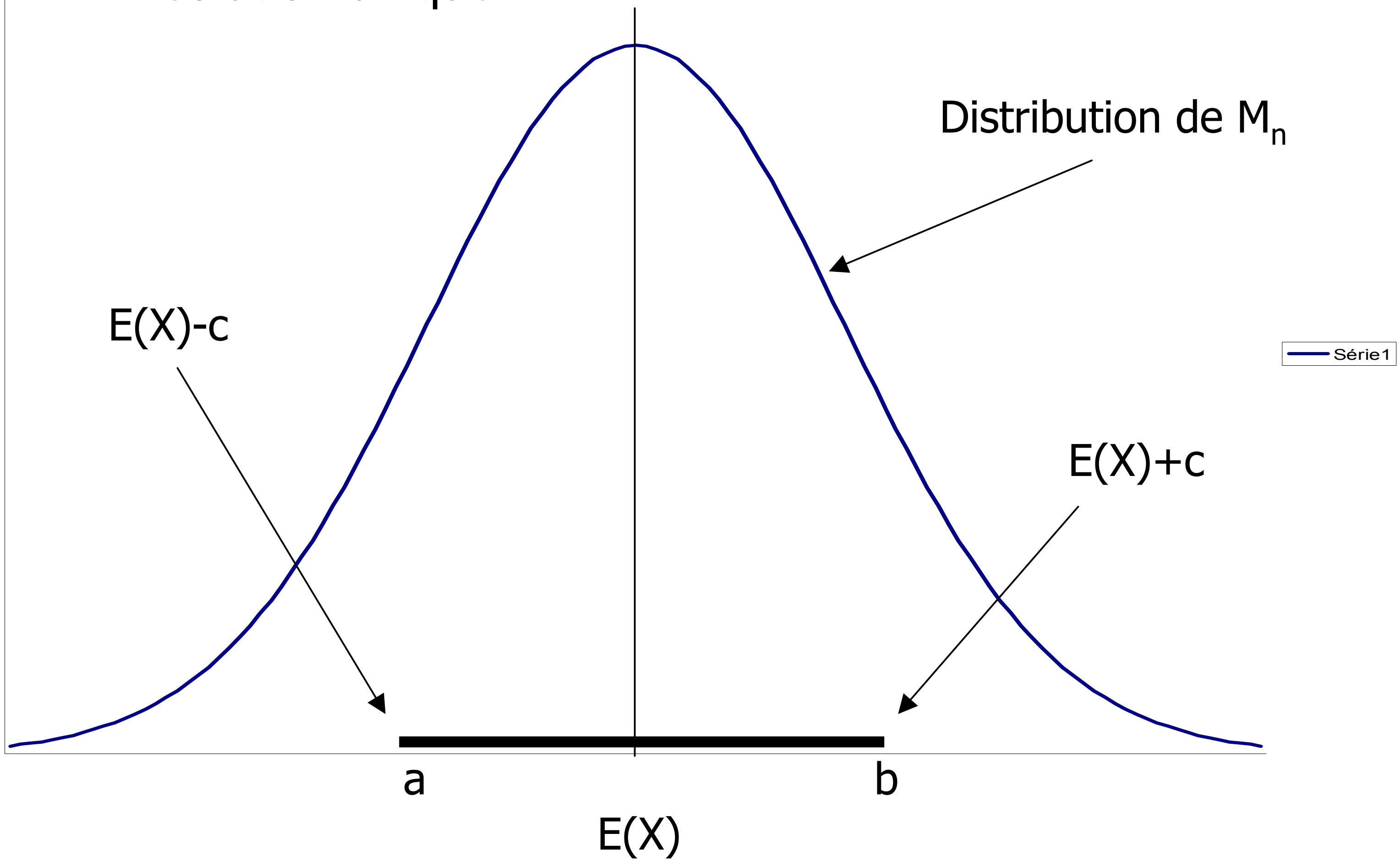


Application du théorème central limite (4)



Application du théorème central limite (5)

- Les solutions sont multiples: choisir [a b] centré en E(X) → solution unique



Application du théorème central limite (6)

- Solution: on suppose que les conditions d'application du théorème central limite s'appliquent

$$M_n \in [E(X) - c \;\; E(X) + c] \Leftrightarrow M_n - E(X) \in [-c \;\; c] \;\; \Leftrightarrow$$

$$\frac{M_n - E(X)}{\frac{\sigma(X)}{\sqrt{n}}} \in [\frac{-c}{\frac{\sigma(X)}{\sqrt{n}}} \;\; \frac{c}{\frac{\sigma(X)}{\sqrt{n}}}]$$

$$P(\frac{M_n - E(X)}{\frac{\sigma(X)}{\sqrt{n}}} \in [\frac{-c}{\frac{\sigma(X)}{\sqrt{n}}} \;\; \frac{c}{\frac{\sigma(X)}{\sqrt{n}}}]) = 1 - \alpha$$

Application du théorème central limite (7)

$$P\left(\frac{M_n - E(X)}{\frac{\sigma(X)}{\sqrt{n}}} \in \left[\frac{-c}{\frac{\sigma(X)}{\sqrt{n}}} \quad \frac{c}{\frac{\sigma(X)}{\sqrt{n}}}\right]\right) = 1 - \alpha$$

$\sim N(0,1)$

$[-u_\alpha \qquad u_\alpha]$

$$IP_{1-\alpha}(M_n) = \left[\, E(X) - u_\alpha \frac{\sigma(X)}{\sqrt{n}} \quad E(X) + u_\alpha \frac{\sigma(X)}{\sqrt{n}} \,\right]$$

$n \geq 30$

- Pour une variable de Bernoulli

$$IP_{1-\alpha}(P_n) = \left[\, \pi - u_\alpha \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \quad \pi + u_\alpha \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \,\right]$$

$n\pi \geq 5$ et $n(1-\pi) \geq 5$

Terminologie

- Généralisation
  Tout intervalle [a b] tel que

$$\Pr(X \in [a\ b]) = 1 - \alpha$$

  se nomme INTERVALLE DE PARI POUR X
  - de niveau $1-\alpha$
  - au risque α

- Site:

http://www.jcu.edu/math/isep/Quincunx/Quincunx.html