Statistiques et informatique

appliquées aux sciences sociales

Traiter des données numériques 1

Mardi 7/11/2023 15h30-17h Censier D2

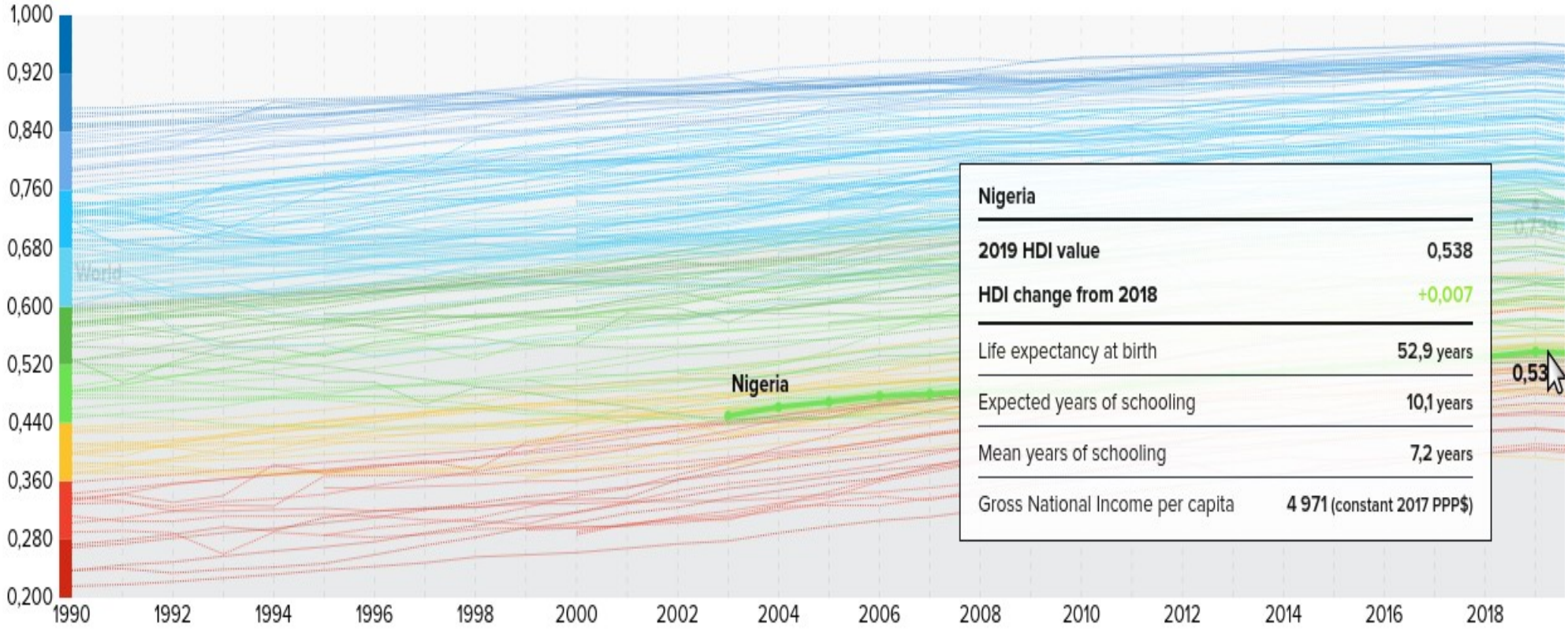
Licence de sciences sociales 3e année
Université Paris 1 Panthéon Sorbonne

Plan de la séance

- Exemple de données numériques
- Statistiques univariées
- Représentations graphiques (1)
- Détermination des quantiles

Variables numériques

- Âge, revenu, budget, consommation, patrimoine mais aussi surface du logement, nombre de pièces, nombre de partenaires et variables idéaltypiques (ex : comptage des modalités représentatives de la bonne volonté scolaire) pour une personne ou un ménage, nombreuses variables pour les entreprises (chiffre d'affaire, valeur ajoutée, profit, nombre de salariés, capital, etc.)
- Données agrégées (comptages, sommes, moyennes, autres indicateurs) :
  PIB, PIB/tête, IDH, etc.
- Type de variable omniprésent en statistiques, auquel on peut toujours se ramener : en effet, une proportion se calcule comme une moyenne à partir d'une variable dichotomique
- Importance de bien définir la Population d'étude et donc les individus statistiques (des pays dans l'exemple qui suit)


1,000
0,920
0,840
0,760
0,680
0,600
0,520
0,440
0,360
0,280
0,200
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
Nigeria
0,53
Nigeria
2019 HDI value 0,538
HDI change from 2018 +0,007
Life expectancy at birth 52,9 years
Expected years of schooling 10,1 years
Mean years of schooling 7,2 years
Gross National Income per capita 4 971 (constant 2017 PPP$)

Exemple : données du PNUD (2017)

Pays	IDH 2017	PIB/tête (2017)	Esp Vie Naiss 2017	Années éducation (2017)	Population totale (2017)	Ratio Palma 2010-17
Bangladesh	0,608	3677	72,8	11,4	164,7	1,3
Brésil	0,759	13755	75,7	15,4	209,3	3,5
Chine	0,752	15270	76,4	13,8	1409,5	2,1
États-Unis	0,924	54941	79,5	16,5	324,5	2
Fédération de Russie	0,816	24233	71,2	15,5	144	1,6
Inde	0,64	6353	68,8	12,3	1339,2	1,5
Indonésie	0,694	10846	69,4	12,8	264	1,8
Japon	0,909	38986	83,9	15,2	127,5	1,2
Nigéria	0,532	5231	53,9	10	190,9	2,2
Pakistan	0,562	5311	66,6	8,6	197	1,2

Indicateurs de tendance centrale

- Objectif : résumer la distribution par une valeur « centrale »
- Moyenne : la plus proche des différentes valeurs au sens des moindres carrés, indicateur « sensible » aux valeurs extrêmes, parfois calculée en enlevant des « points aberrants » (en raison de sa sensibilité)
- Médiane : divise la distribution en deux moitiés d’effectifs égaux, indicateur « robuste »
- Mode : intéressant si des valeurs identiques se retrouvent en grand nombre
- Moyenne géométrique, logarithmes, utilisée par exemple sur des échelles de revenu

Indicateurs de dispersion

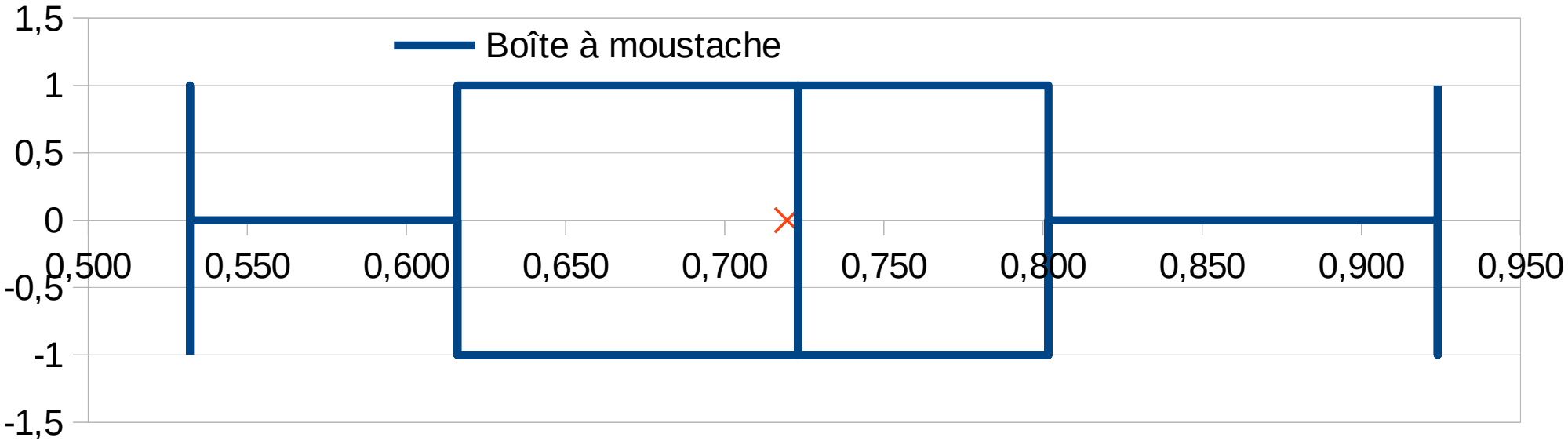
- Objectif : mesurer les écarts entre valeurs
- Étendue de la distribution : Max-Min
- Écart-type : racine carrée de la variance théorique ou empirique, moyenne des carrés des écarts à la moyenne
- Écart interquartiles, différence entre le valeur la moins élevée du quart supérieur et la valeur la plus élevée du quart inférieur
- Autres quantiles, déciles, centiles…
- Mesures relatives : Coefficient de variation, écart interquartile relatif, rapport interquartiles ou interdéciles

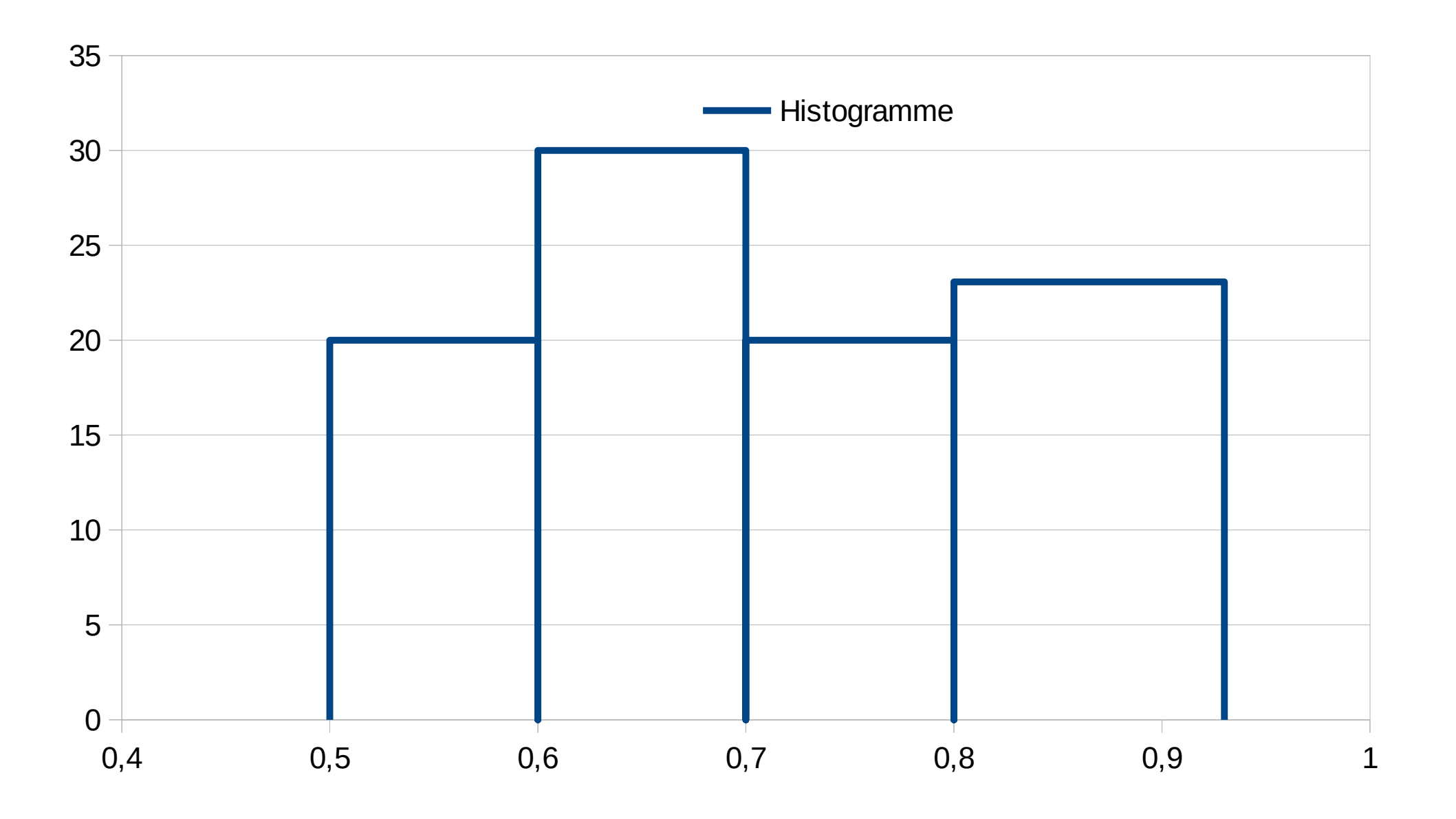
Indicateur	IDH 2017	PIB/tête (2017)	Esp Vie 2017	Ann édu (2017)	Population totale (2017)	Ratio Palma
Somme	7,196	178603	718,2	131,5	4370,6	18,4
Moyenne par pays	0,7196	17860,3	71,82	13,15	437,06	1,84
Moyenne par hab.	*0,7104*	*14806,3*	*72,33*	*13,16*	*pas de sens*	*1,85*
Médiane	0,723	12300,5	72	13,3	203,15	1,7
Q1	0,616	5571,5	68,95	11,625	171,25	1,35
Q3	0,80175	21992,25	76,225	15,35	309,375	2,075
Variance	0,0169	258521647	60,5636	6,0965	222810,5	0,4264
Écart-type	0,130	16079	7,782	2,469	472,028	0,653
Coeff. de variation	0,181	0,900	0,108	0,188	1,080	0,355
Q3-Q1	0,18575	16420,75	7,275	3,725	138,125	0,725
relatif	0,257	1,335	0,101	0,280	0,680	0,426
Q3/Q1	1,3015	3,948	1,106	1,320	1,807	1,537
D9/D1	1,629	7,995	1,224	1,582	9,457	1,942

Représentations d’une distribution

- Visualiser les quartiles et la moyenne avec une « boîte à moustache »
- Résumer la distribution à l’aide de tranches grâce un histogramme
- Déterminer n’importe quel quantile avec un diagramme cumulé croissant

Représentation d’une distribution




35
30
25
20
15
10
5
0
0,4
0,5
0,6
0,7
0,8
0,9
1
Histogramme

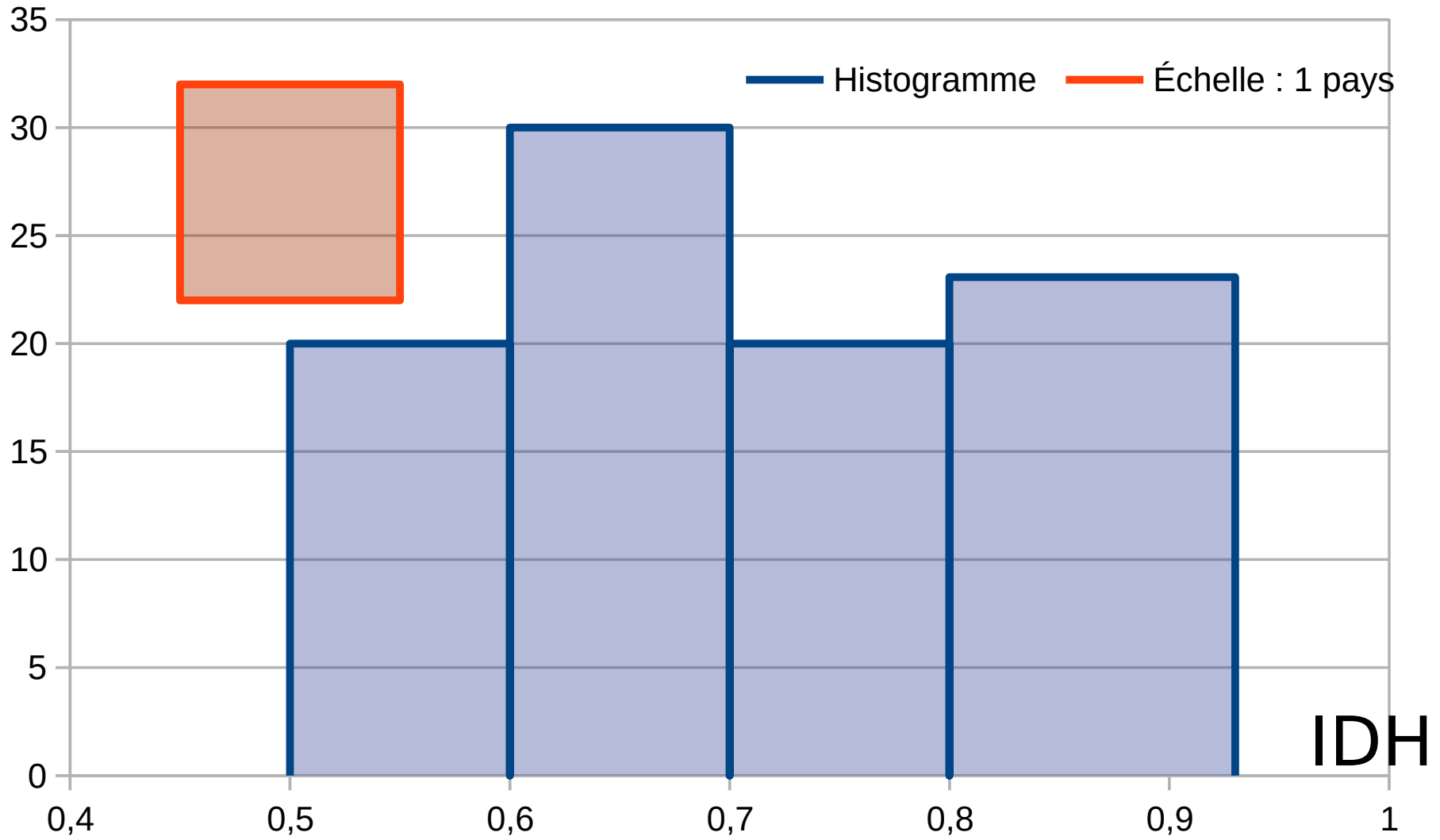
Principe d’un histogramme

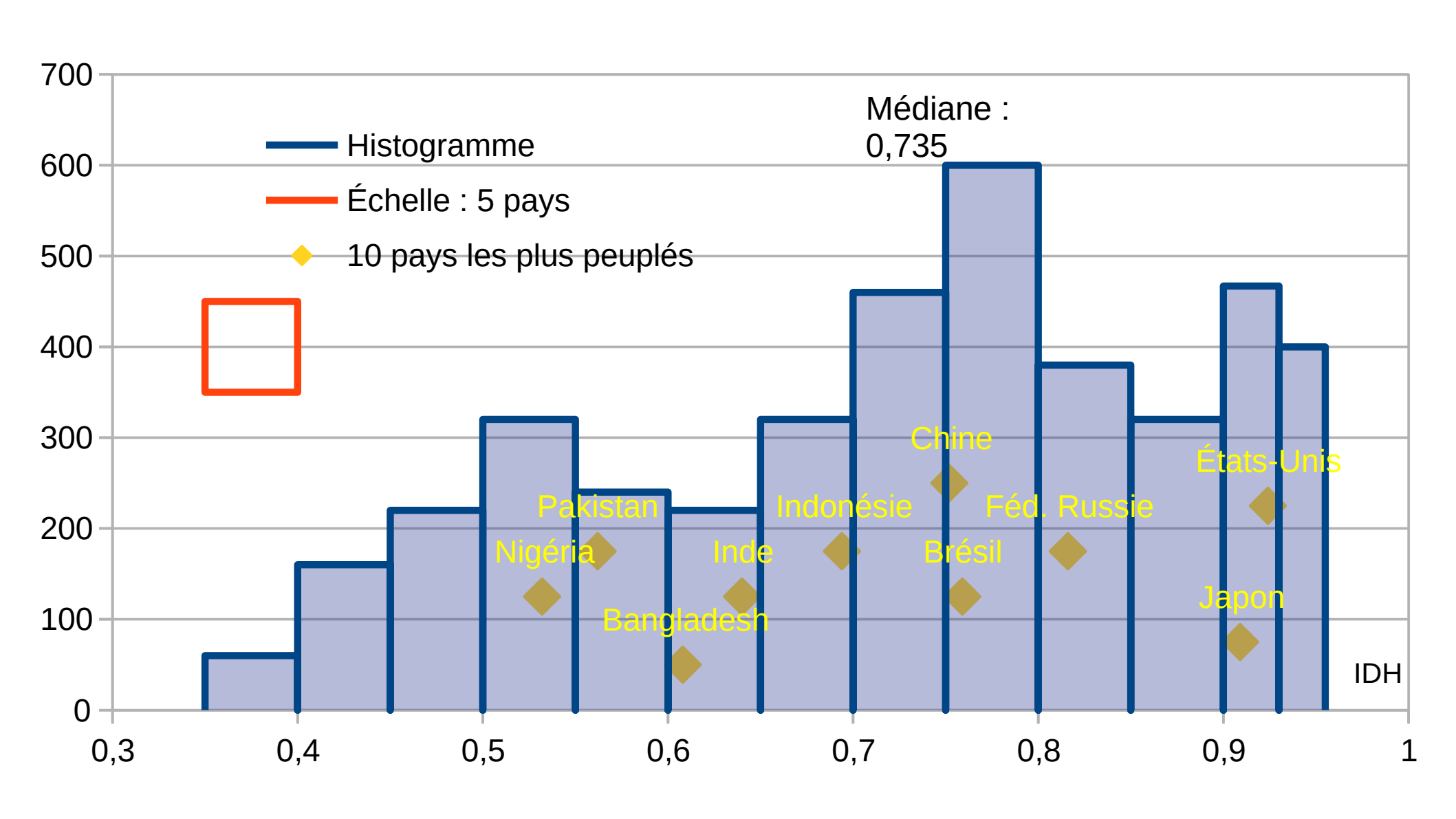
- L’histogramme est une représentation en surface d’une distribution en classes
- Il faut donc au préalable regrouper les valeurs de la distribution dans des intervalles : certaines données peuvent être d’emblée sous cette forme (par exemple, on demande des tranches de revenu au lieu de demander le revenu exact)
- Pour des données suffisamment nombreuses (ex : IDH des 197 pays recensés par le PNUD), on peut regrouper systématiquement avec un intervalle de largeur fixe

Pays	IDH 2017	Effectif cumulé	Classes
Nigéria	0,532	1	[0,5 ; 0,6 [
Pakistan	0,562	2	[0,5 ; 0,6 [
Bangladesh	**0,608**	**3**	[0,6 ; 0,7 [
Inde	0,64	4	[0,6 ; 0,7 [
Indonésie	0,694	5	[0,6 ; 0,7 [
Chine	0,752	6	[0,7 ; 0,8 [
Brésil	0,759	7	[0,7 ; 0,8 [
Féd. Russie	**0,816**	**8**	[0,8 ; 0,93 [
Japon	0,909	9	[0,8 ; 0,93 [
États-Unis	0,924	10	[0,8 ; 0,93 [

Classes	Amplitude	Effectif	Densité
[0,5 ; 0,6 [	0,1	2	20
[0,6 ; 0,7 [	0,1	3	30
[0,7 ; 0,8 [	0,1	2	20
[0,8 ; 0,93 [	0,13	3	23,1

Indicateur	IDH 2017	PIB/tête (2017)	Esp Vie 2017	Ann édu (2017)	Population totale (2017)	Ratio Palma
Somme	7,196	178603	718,2	131,5	4370,6	18,4
Moyenne par pays	0,7196	17860,3	71,82	13,15	437,06	1,84
Moyenne par hab.	*0,7104*	*14806,3*	*72,33*	*13,16*	*pas de sens*	*1,85*
Médiane	0,723	12300,5	72	13,3	203,15	1,7
Q1	0,616	5571,5	68,95	11,625	171,25	1,35
Q3	0,80175	21992,25	76,225	15,35	309,375	2,075
Variance	0,0169	258521647	60,5636	6,0965	222810,5	0,4264
Écart-type	0,130	16079	7,782	2,469	472,028	0,653
Coeff. de variation	0,181	0,900	0,108	0,188	1,080	0,355
Q3-Q1	0,18575	16420,75	7,275	3,725	138,125	0,725
relatif	0,257	1,335	0,101	0,280	0,680	0,426
Q3/Q1	1,3015	3,948	1,106	1,320	1,807	1,537
D9/D1	1,629	7,995	1,224	1,582	9,457	1,942


35
30
25
20
15
10
5
0
0,4
0,5
0,6
0,7
0,8
0,9
1
Histogramme
Échelle : 1 pays
IDH


700
600
500
400
300
200
100
0
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
Médiane :
0,735
Histogramme
Échelle : 5 pays
10 pays les plus peuplés
Chine
États-Unis
Pakistan
Indonésie
Féd. Russie
Nigéria
Inde
Brésil
Japon
Bangladesh
IDH

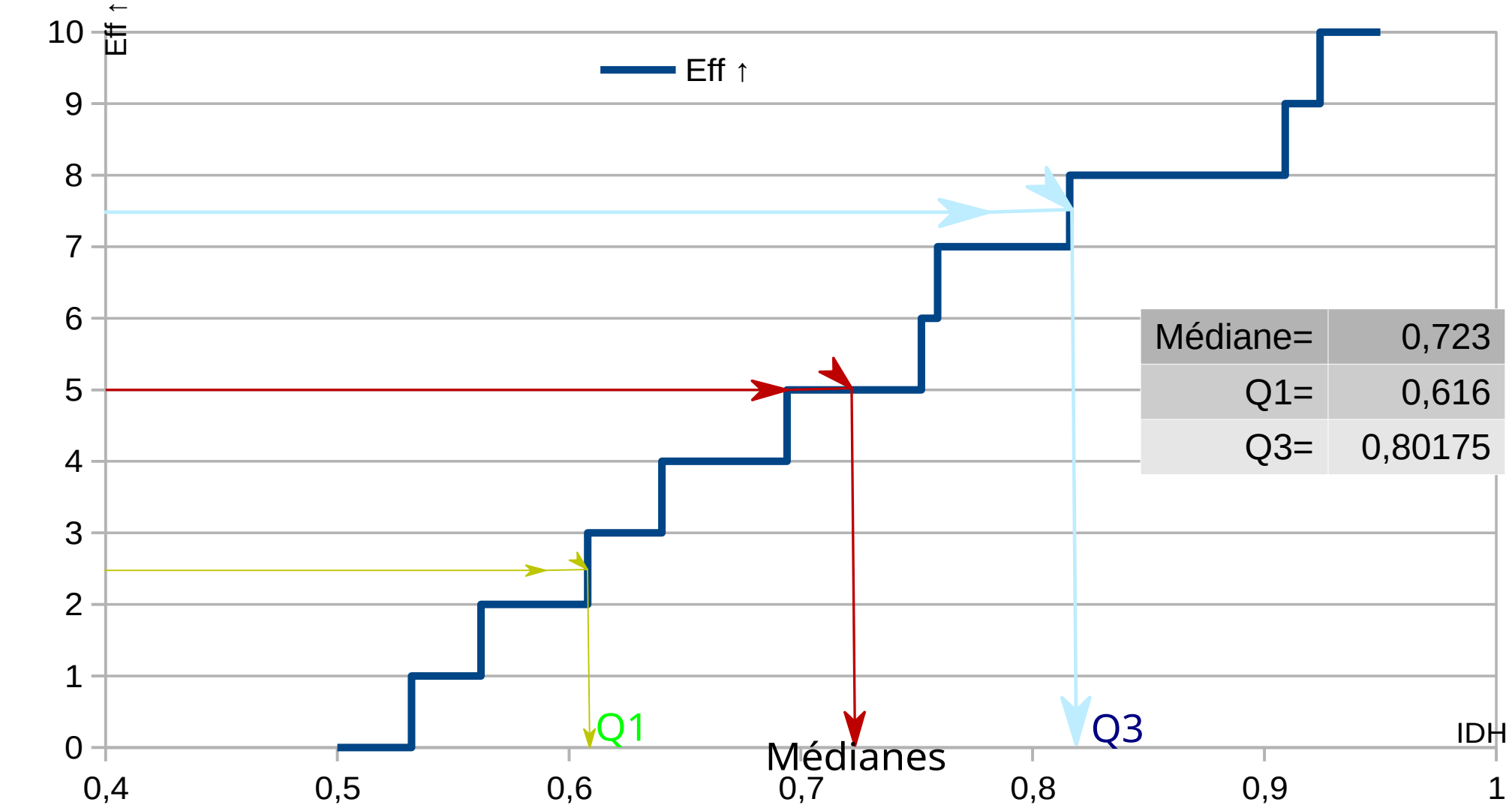
Détermination de la médiane

- N=10 : n/2=5 donc toute valeur située entre le 5^{e} (Indonésie, IDH=0,694) et le 6^{e} pays (Chine, IDH=0,752) ordonnés par IDH convient. On peut prendre la moyenne de ces deux valeurs soit 0,723
- On procède de même pour les autres quantiles. Par exemple, pour Q1 le premier quartile, n/4=2,5, on prend l'IDH du Bangladesh, 3^{e} sur la liste ordonnée car 1/4 de la distribution est en dessous de son IDH et 3/4 au-dessus en incluant à chaque fois ce pays donc Q1=0,608

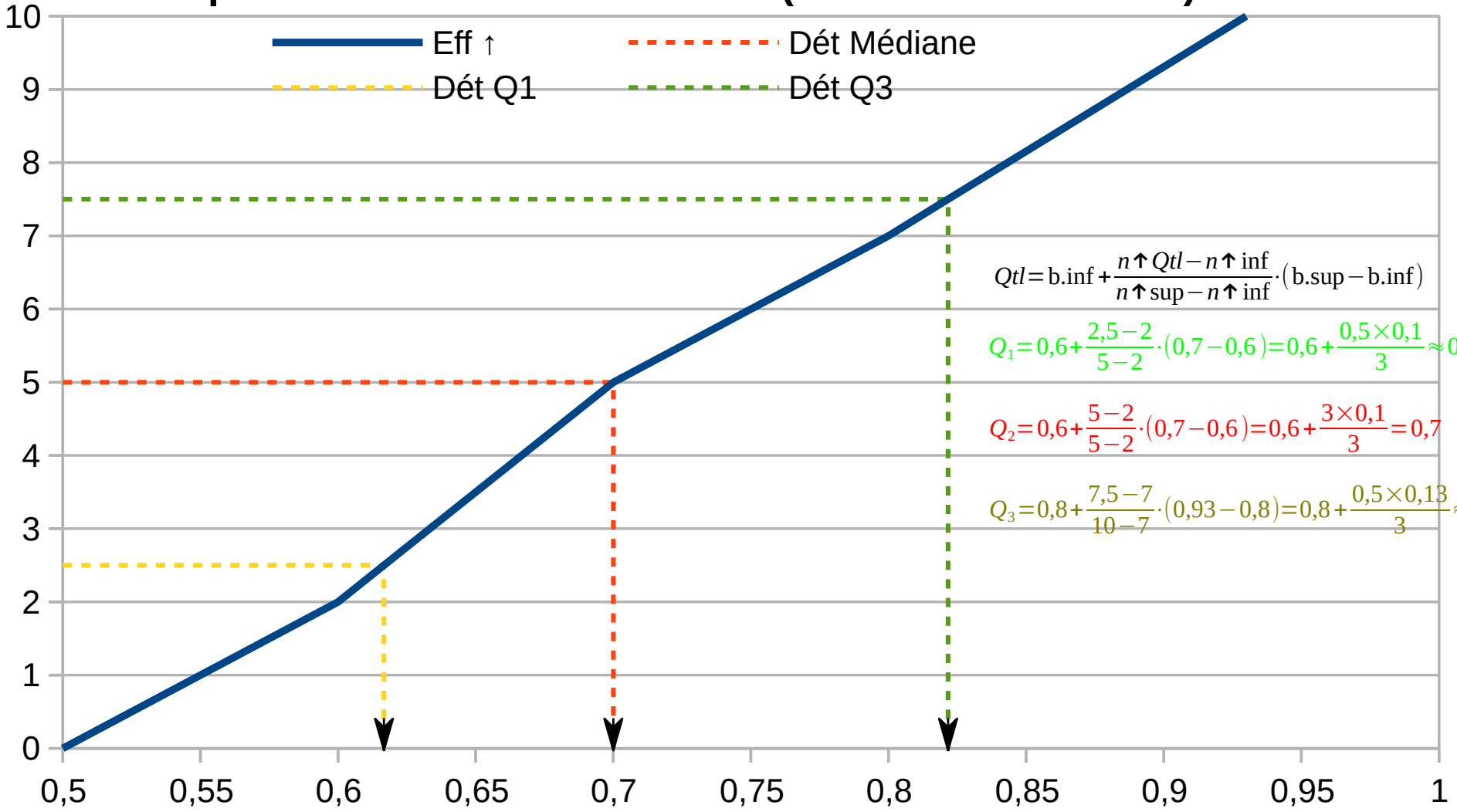
Détermination géométriques des quantiles

- Avec les données d’origine
- Avec les données en classes (interpolation)

Courbe de répartition cumulée croissante (effectif inférieur ou égal à une valeur)



Courbe de répartition cumulée croissante (données en classes)



$$Qtl = \text{b.inf} + \frac{n\uparrow Qtl - n\uparrow \text{inf}}{n\uparrow \text{sup} - n\uparrow \text{inf}} \cdot (\text{b.sup} - \text{b.inf})$$

$$Q_1 = 0{,}6 + \frac{2{,}5-2}{5-2} \cdot (0{,}7-0{,}6) = 0{,}6 + \frac{0{,}5\times 0{,}1}{3} \approx 0{,}617$$

$$Q_2 = 0{,}6 + \frac{5-2}{5-2} \cdot (0{,}7-0{,}6) = 0{,}6 + \frac{3\times 0{,}1}{3} = 0{,}7$$

$$Q_3 = 0{,}8 + \frac{7{,}5-7}{10-7} \cdot (0{,}93-0{,}8) = 0{,}8 + \frac{0{,}5\times 0{,}13}{3} \approx 0{,}822$$