

Statistique

Introduction

- I. Statistique descriptive
- II. Bases en probabilités
- III. Statistique probabiliste

Introduction

1. Préambule
2. Terminologie
3. Un art de mentir ?
4. Echantillonnage

1. Préambule

Question / collectif

Etat quantitatif = *une* statistique

Méthodes = *la* Statistique

réaliser l'état

le présenter (Stat. descriptive)

l'analyser

2. Terminologie

Le collectif = Population

Un élément = Individu

Une partie = Echantillon

Variable Qualitative

Quantitative Discrète

Continue

Une valeur = Modalité

3. Un art de mentir ?

Réponse scientifique

Rigueur, honnêteté, clarté,
serrer le réel

Causes de défiance

Questions sensibles

biais volontaire

Echantillonnage

limites de principe

4. Echantillonnage

Population trop vaste échantillonnage

Echantillonnage écart aléatoire de réponse

Moindre écart représentativité

Représentativité équité du choix

= « randomisation »

Equité analyse probabiliste

Equité impraticable sondage ; aléa biaisé

I. Statistique descriptive

1. Indicateurs numériques
2. Tables
3. Graphiques

I.1. Indicateurs numériques

Variable	V
Nombre de modalités	M
Modalité de V indicée	j (1 à M)
Taille d'échantillon	N
Individu indicé i (1 à N)	
Valeur de V pour l'individu i	V_i
Variable <u>continue</u> discrétisée en classes $[v_{inf_j} ; v_{sup_j}]$	
Une classe = une modalité d'expression	$V_j = (v_{sup_j} + v_{inf_j}) / 2$

I.1.a Effectif, fréquence

Nombre d'individus à modalité v_j n_j
= **Effectif** (de la modalité v_j de V)

Fraction d'individus $f_j = n_j / N$
= **Fréquence** (de la modalité v_j de V)

Variable qualitative ou quantitative

I.1.b Moyenne (d'échantillon)

Mesure de tendance centrale

$$m_e = \frac{1}{N} \sum_{i=1}^N v_i$$

Changement d'unité

v_i devient $a v_i + b$

m_e devient $a m_e + b$

Table fréquence/valeur (f_j, v_j)

$$m_e = \sum_{j=1}^M f_j v_j$$

I.1.c.α Variance (d'échantillon)

Mesure de dispersion

$$s_e^2 = \frac{1}{N} \sum_{i=1}^N (v_i - m_e)^2$$

Changement d'unité

v_i	devient	$a v_i + b$
s_e^2	devient	$a^2 s_e^2$

Table fréquence/valeur (f_j , v_j)

$$s_e^2 = \sum_{j=1}^M f_j (v_j - m_e)^2 = \left(\sum_{j=1}^M f_j v_j^2 \right) - m_e^2$$

I.1.c.β Ecart-type (d'échantillon)

Variable V	unité U	(kg)	(m)
Variance s_e^2	unité U^2	(kg ²)	(m ²)

Ecart-type

$$s_e = \sqrt{s_e^2}$$

Il a même unité que V

V / s_e est sans unité

I.1.d Médiane

Ordonner les v_i

N pair $k=N/2$ $C_{50} = (v_k + v_{k+1}) / 2$

N impair $k=(N-1) / 2$ $C_{50} = v_{k+1}$

$$C_{50} = \text{« centile 50 »} = \text{médiane}$$

Insensible aux valeurs extrêmes (usage en biologie)

I.1.e Mode

Modalité v_j d'un maximum (local) de fréquence
= mode

Plusieurs sommets

Autant de modes

= distribution multimodale

(bimodale, trimodale...)

I.1.f.α Covariance (d'échantillon)

Mesure de liaison entre 2 variables X et Y

$$\text{COV}_{X\text{Ye}} = \frac{1}{N} \sum_{i=1}^N (x_i - m_{xe})(y_i - m_{ye})$$

Pour le calcul :

$$\text{COV}_{X\text{Ye}} = \left(\frac{1}{N} \sum_{i=1}^N x_i y_i \right) - m_{xe} m_{ye}$$

I.1.f.β Corrélation (d'échantillon)

Mesure de liaison entre variables X et Y

$$\text{corr}_{XYe} = \frac{\text{COV}_{XYe}}{S_{Xe} S_{Ye}}$$

Sans unité

(sans dimension, indépendant du choix d'unité)

I.1.g Autres indicateurs

Etendue (mesure de dispersion)

$$\text{Etendue} = \max(v_i) - \min(v_i)$$

Ecart moyen en valeur absolue

$$\text{Emva} = \frac{1}{N} \sum_{i=1}^N |v_i - m_e|$$

I.2. Tables

Répartition des valeurs de V (distribution des v_i)

	information morphologique
Indicateur numérique	Faible
Table	Elevée

Note (terminologie alternative) :

Fréquence = fréquence relative

Effectif = fréquence absolue

I.2.a Tables de fréquences simples

Exemple (données fictives)

Effectif échantillon : $N = 115$

Hospitalisation (jours)	≤ 1	2 à 3	> 3
Effectifs	72	29	14
Fréquences	0,626	0,252	0,122
Fréquences (%)	62,6	25,2	12,2

I.2.b Tables de fréquences cumulées

Exemple (données fictives)

Hospitalisation (jours)	≤ 1	2 à 3	> 3
Effectifs	72	29	14
Cumuls			
Croissant	72	101	115
Décroissant	115	43	14

I.2.c Tables croisées (de contingence)

Exemple (données fictives) N = 700.

Diag.	Taux psa	≤ seuil	> seuil	
hbp		253	38	291
Cp		3	24	27
sain		328	54	382
		584	116	700

I.3. Graphiques

Propriétés informatives $\approx$ tables

Ajoutent : appréciation visuelle
instantanée

Qualités

Sobriété perception pertinente

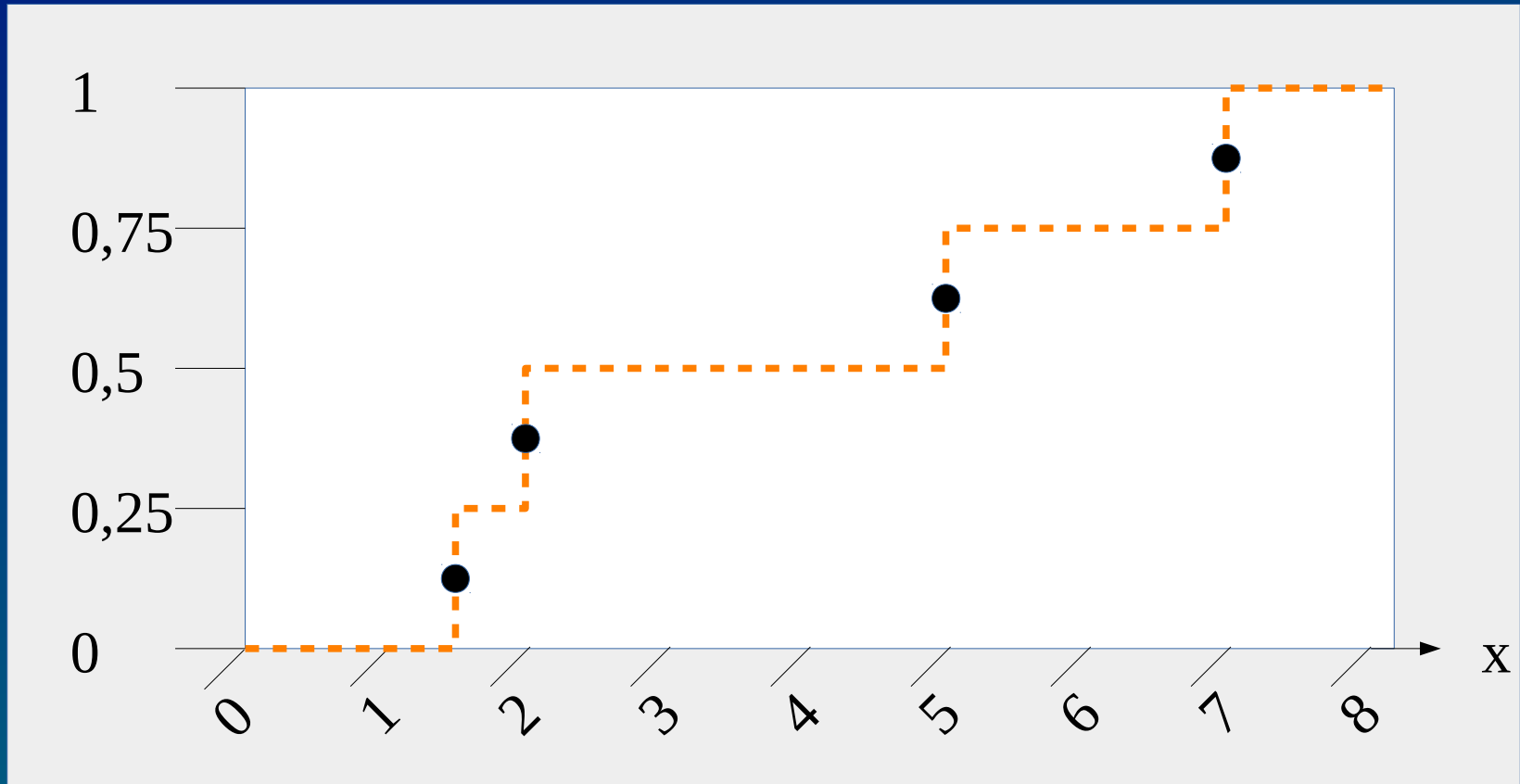
Défauts

Mal exécutés fort biais de perception

I.3.a Graphique des centiles

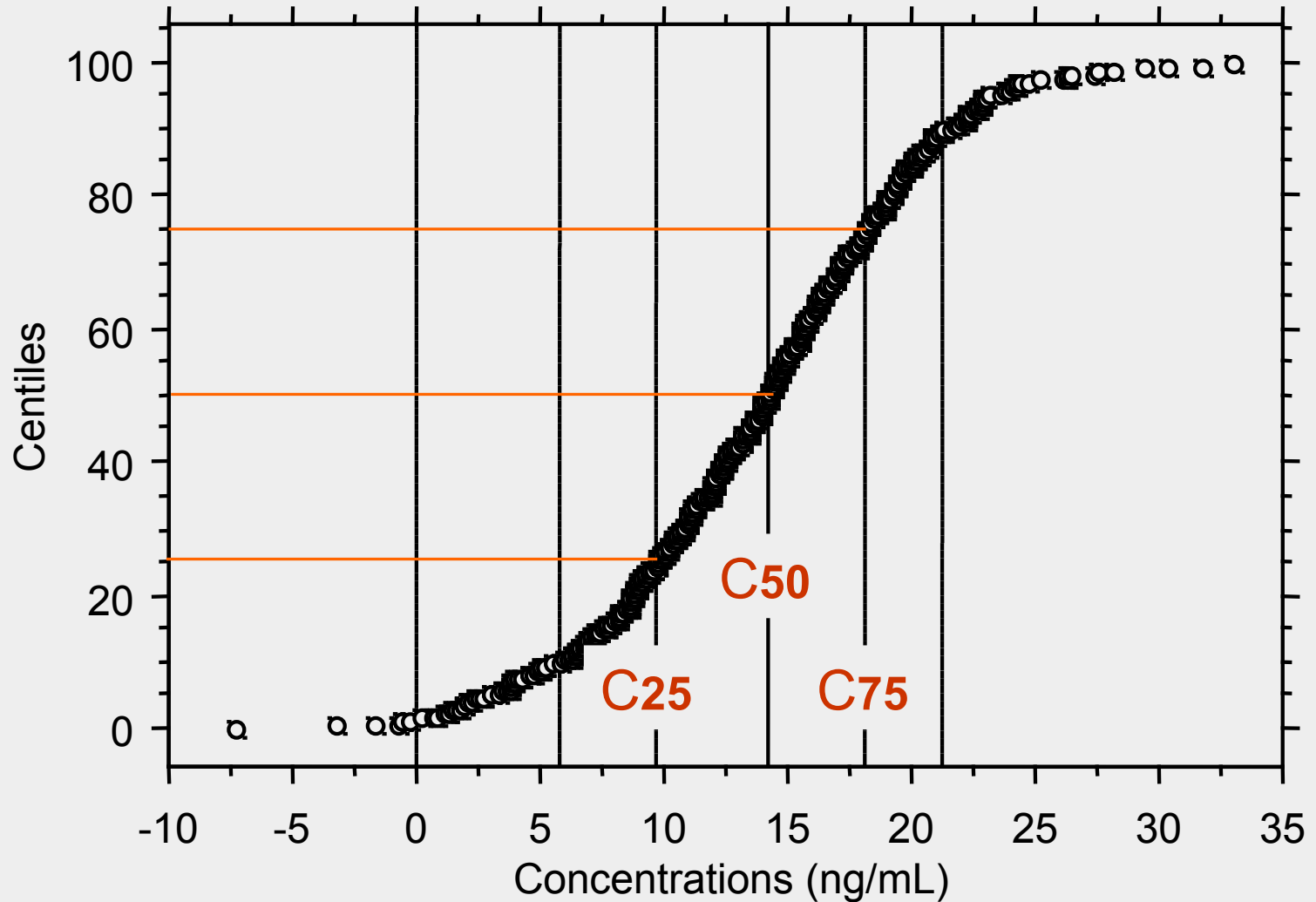
Exemple de construction sur $N = 4$ points

$$(v_i) = (1,5 ; 2 ; 5 ; 7)$$



I.3.a Graphique des centiles (suite)

Un vrai



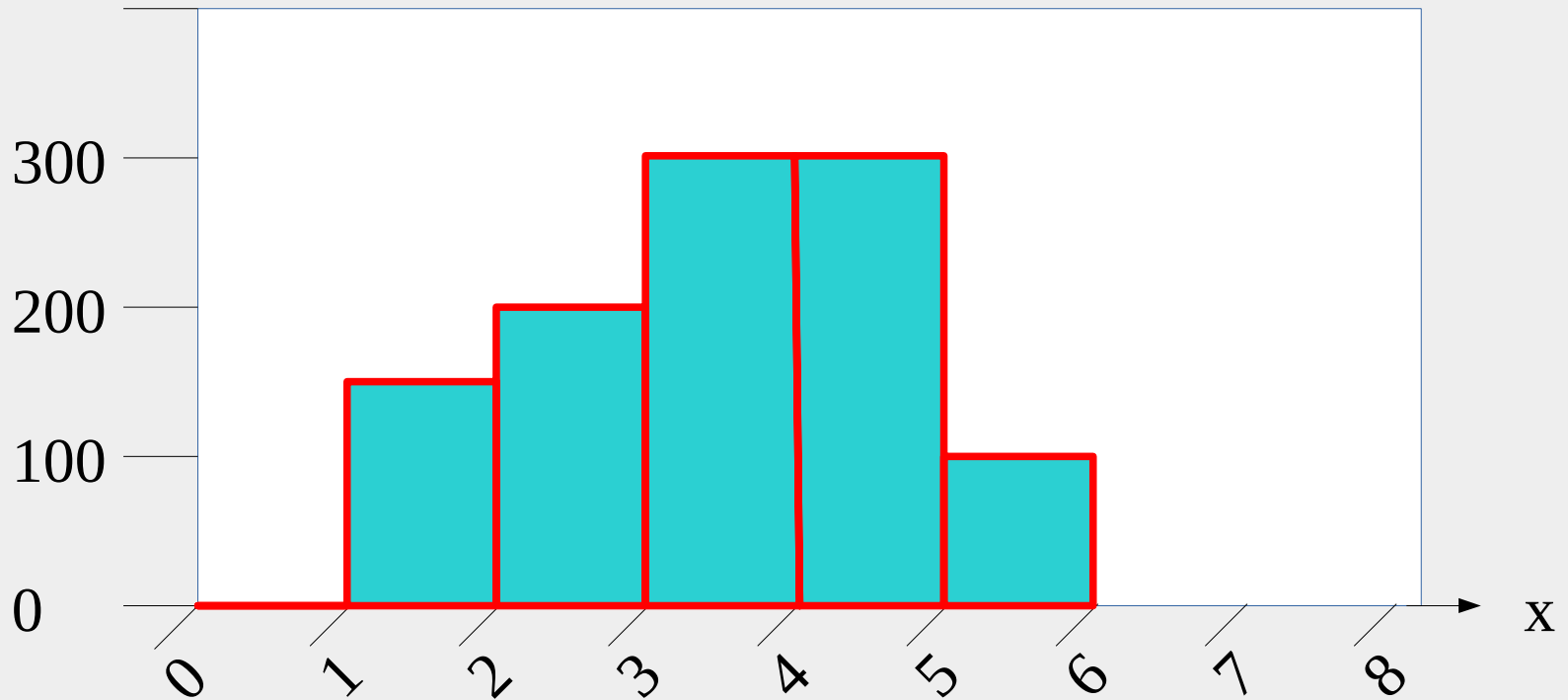
C_{25} = Quartile Q1

C_{75} = Quartile Q3

I.3.b Histogramme

Exemple

effectif/unité de x



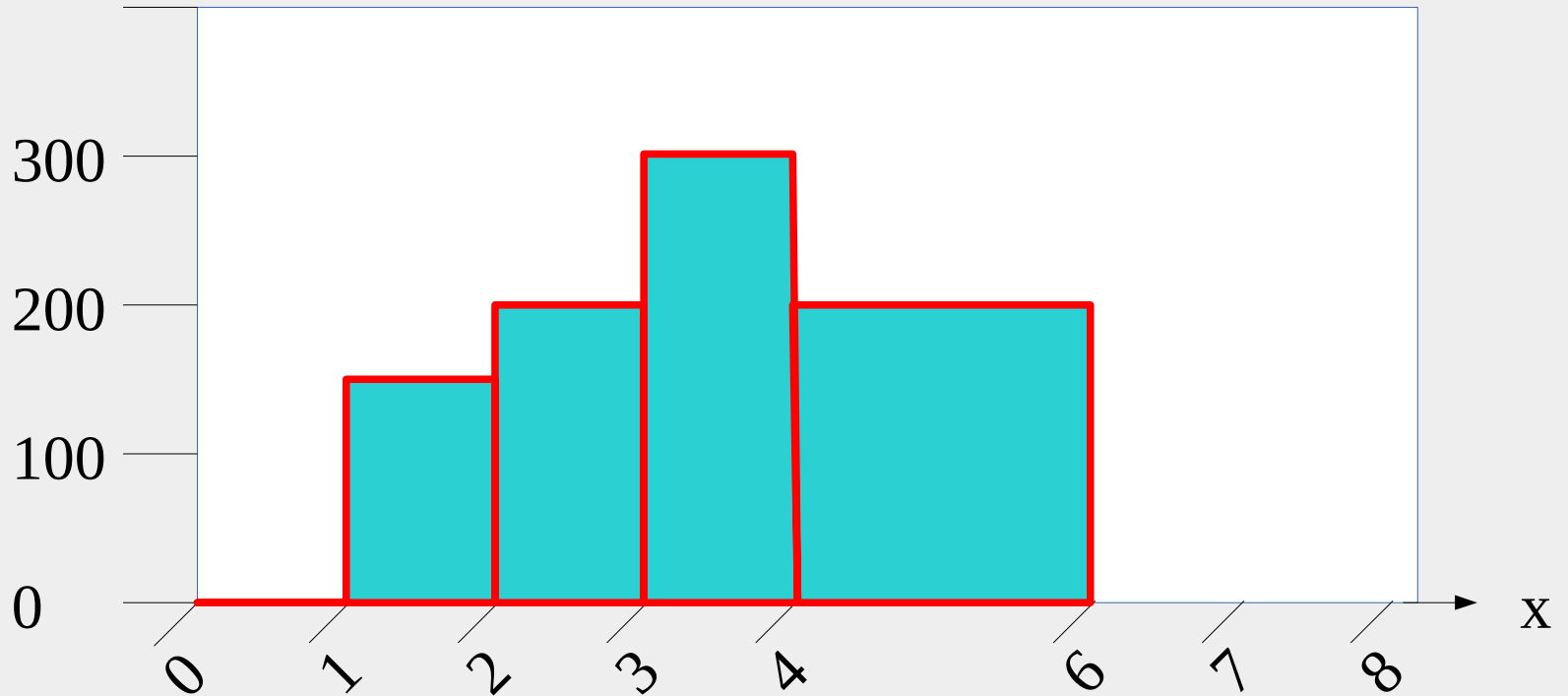
Ici, classes $[j; j+1[$

Surfaces proportionnelles aux effectifs

I.3.b Histogramme

Même exemple avec classe $[4;6[$

effectif/unité de x

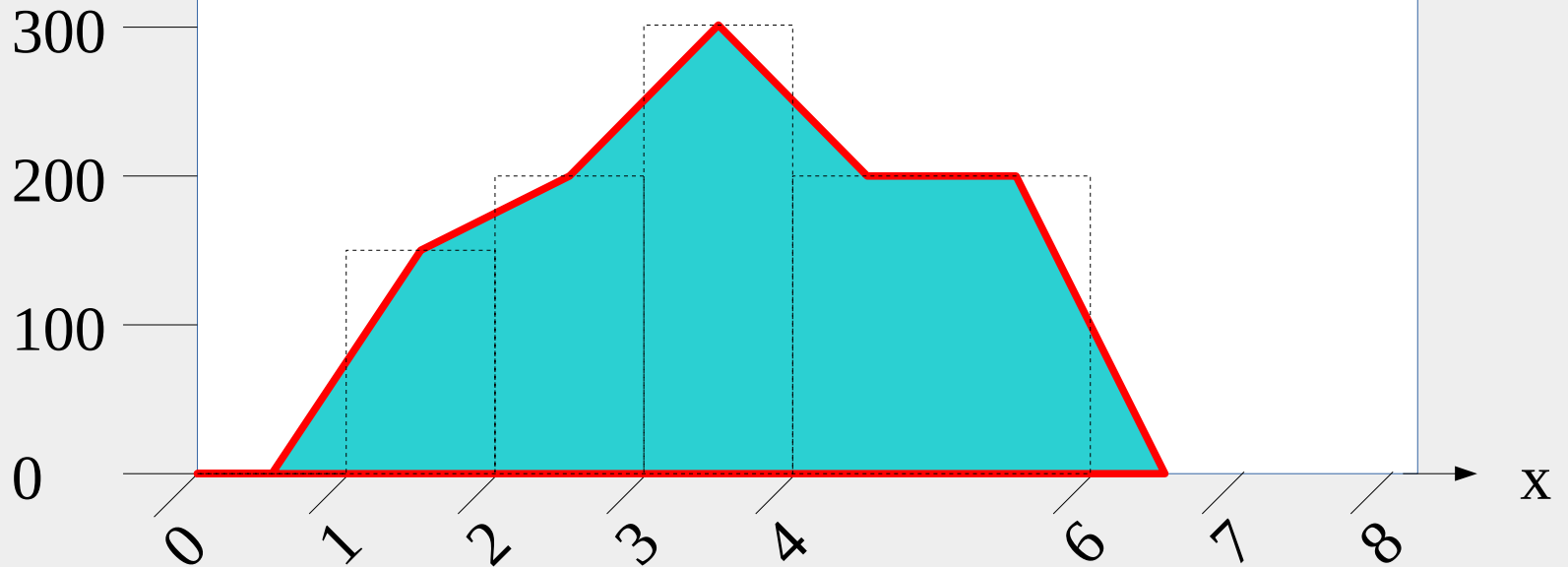


Surfaces toujours proportionnelles aux effectifs

I.3.c Polygone de fréquences

Exemple

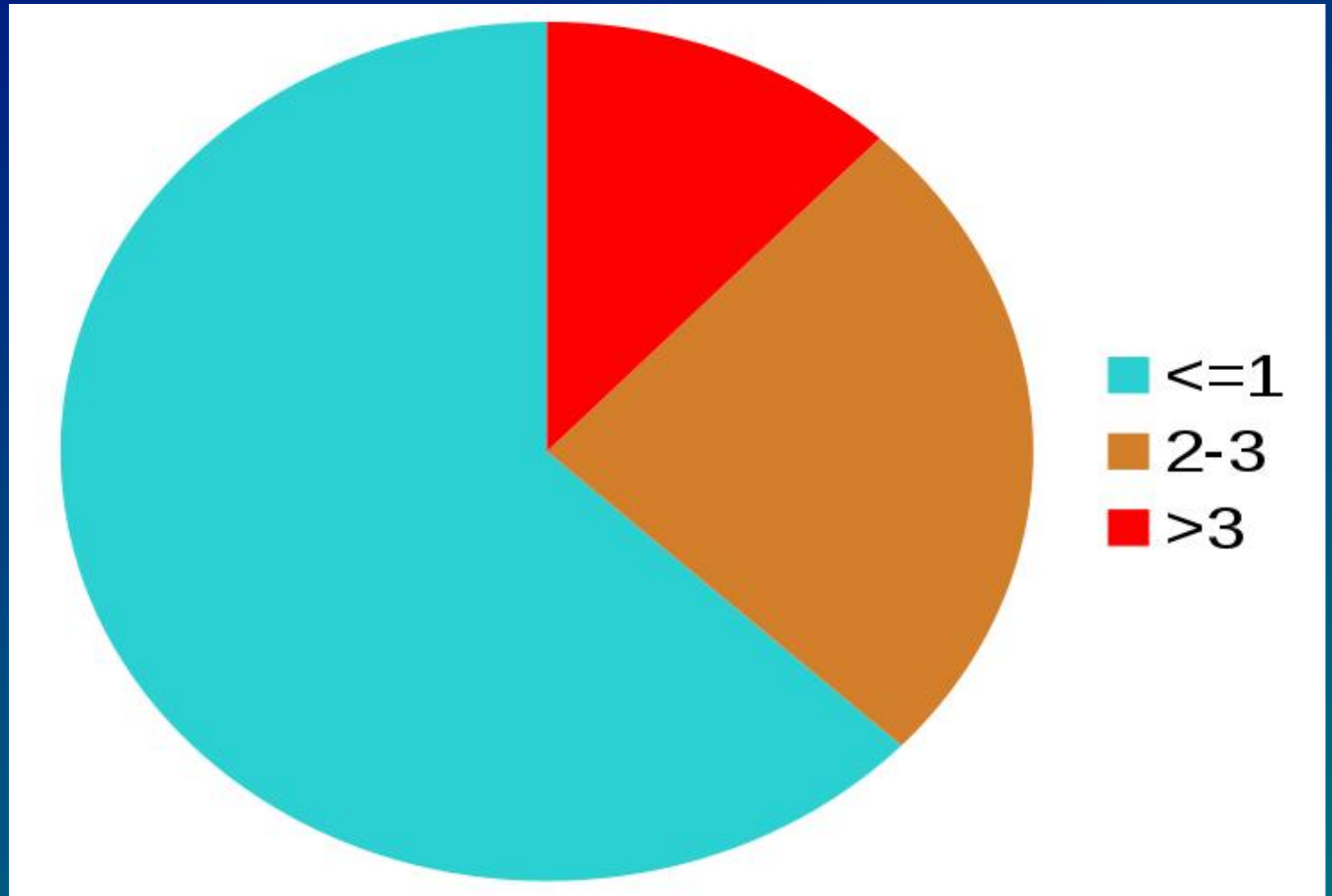
effectif/unité de x



Surfaces proportionnelles conservées

I.3.d Graphe circulaire

Exemple
(voir I.2.a)

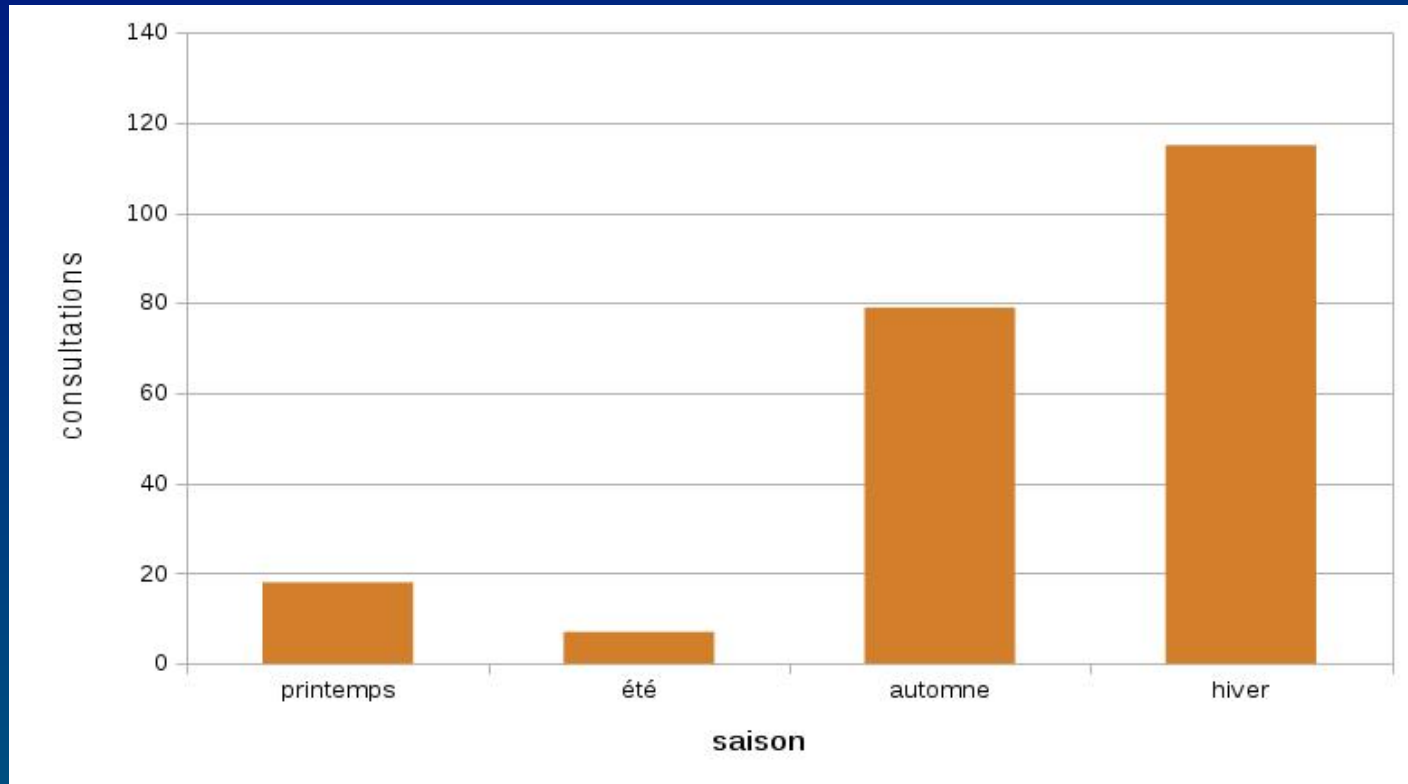


Secteurs proportionnels aux fréquences

I.3.d Graphe en barres

Exemple

(données fictives)

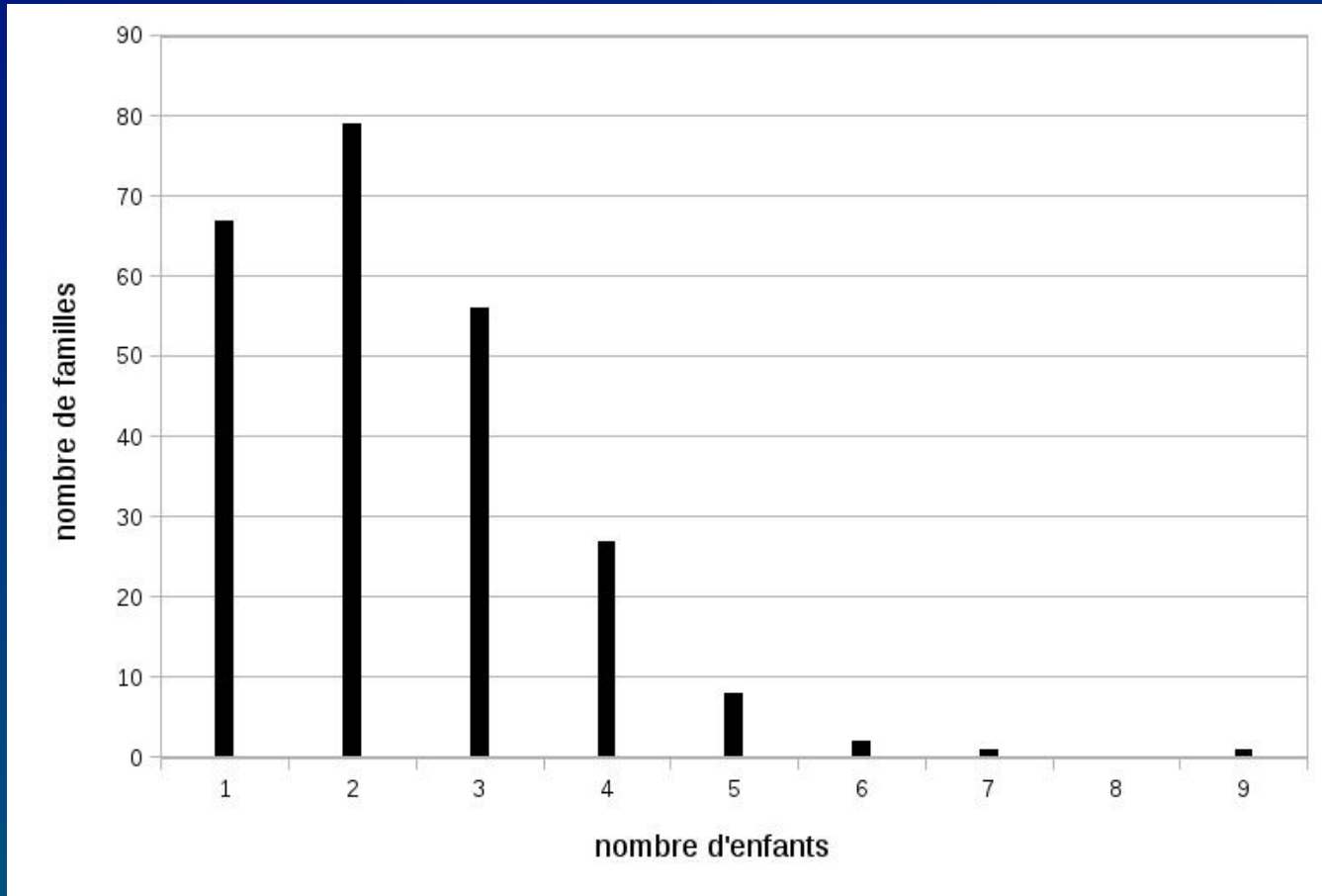


Variable qualitative

Barres : surface sans importance

I.3.e Graphe en bâtons

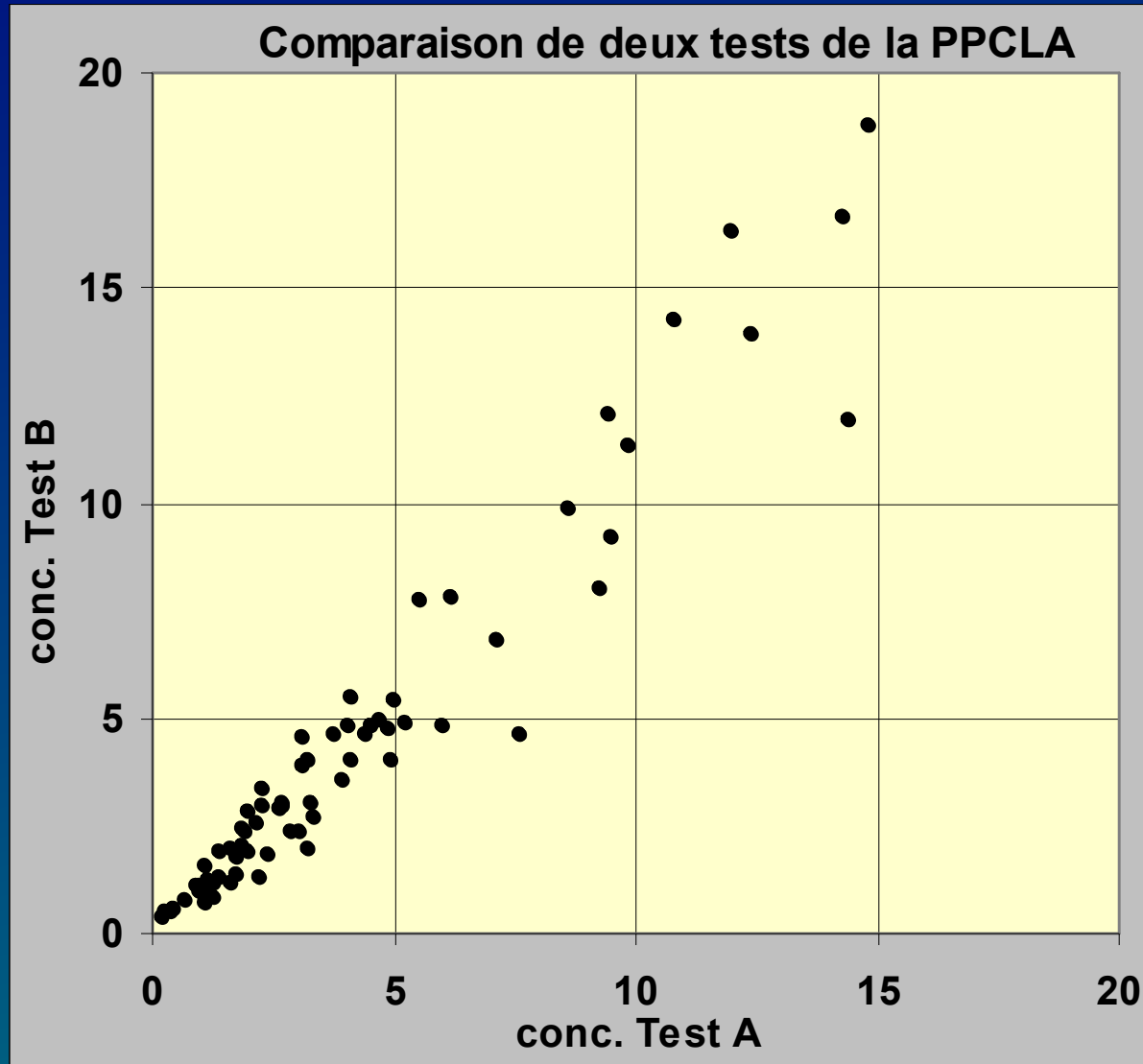
Exemple (données fictives)



Variable quantitative discrète

I.3.f Graphe cartésien (XY)

Exemple (données fictives) N = 70



2 variables quantitatives

I.3.g Autres

Il existe nombre de variantes
des graphes précédents :

- barres horizontales
- barres juxtaposées

Et quantité d'autres graphes

- courbes

- graphes polaires

- à boîtes et moustaches

- en violons

Tous à usage plus spécifique