

Nature des variables et description mono-variable

M8 – Chapitre 1

I. Nature et description des variables

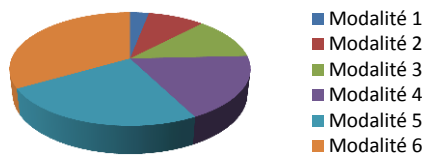
- **Quantitative** : Modalités comparables entre elles
 - **Discrète / continue** : modalités dénombrable / indénombrables
- **Qualitative**
 - **Binaire / Multimodale** : deux modalités / plus de 2.
 - **Ordinale (ou non)** : existence d'un ordre
- **Population** : Ensemble d'évènements possibles
- **Echantillon** : Sous-ensemble de la population
- **Domaine** : Ensemble des valeurs que peut prendre la variable

Définition				Formule		Qualitative binaire	Qualitative multimodale	Qualitative ordinale	Quantitative discrète	Quantitative continue
• Modalités : Ensemble des valeurs distinctes que peut prendre la variable				m_i		X	X	X	X	X
• Effectif n_i : Nombre d'apparitions d'une modalité				n_i		X	X	X	X	X
• Effectif cumulé N_i				$N_i = \sum_{j, m_j \leq m_i} n_j$					X	X
• Fréquence $\hat{f}_i$				$\hat{f}_i = \frac{n_i}{n}$		X	X	X	X	X
• Probabilité				$\mathbb{P}(x = m_i) = \lim_{n \rightarrow +\infty} \hat{f}_i$		X	X	X	X	X
• Fréquence cumulée $\hat{F}_i$				$\hat{F}_i = \sum_{j, m_j \leq m_i} \hat{f}_j$					X	X
• Fonction de répartition empirique $\widehat{F}_X$ (cas discret)				$\widehat{F}_X(x_i) = \widehat{\mathbb{P}}(X \leq m_i) = \hat{F}_i$					X	
• Fonction de répartition empirique $\widehat{F}_X$ (cas continu)				$\widehat{F}_X(x) = \widehat{F}_i^c + (\widehat{F}_i^c - \widehat{F}_{i-1}^c) \frac{x - x_{i-1}}{x_i - x_{i-1}}$ $\widehat{F}_i^c = \widehat{F}_i - \frac{1}{2}(\widehat{F}_i - \widehat{F}_{i-1})$						X
• Fonction de répartition F_X				$F_X(m_i) = \mathbb{P}(X \leq m_i) = \lim_{n \rightarrow +\infty} \hat{F}_i$					X	X
• Moyenne $\bar{x}$	$\min_{\bar{x} \in \mathbb{R}} \sum_{i=1}^n (x_i - \bar{x})^2$	Moyenne empirique : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \hat{f}_i x_i$		Espérance : $\mathbb{E}(X) = \lim_{n \rightarrow \infty} \bar{x}$			x	x	X	X
• Médiane M	$\min_{M \in \mathbb{R}} \sum_{i=1}^n x_i - M $	$\widehat{F}_X(M) = \frac{1}{2}$		$\mathbb{P}(X \leq M) = \frac{1}{2}$		x			X	X
• Mode		$\operatorname{argmax}_{i \in [1; n]} f_i$		$\operatorname{argmax}_{x \in \Omega} \mathbb{P}(x)$		x			X	
• Variance S	$\hat{S}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$	$\hat{s} = \sqrt{\hat{S}^2}$ Ecart type		$S^2 = \mathbb{E}(X - \mathbb{E}^2(X))$			x	x	X	X
• Usage de la variance	$\mathbb{P}(X - \bar{x} \geq k\sigma) \geq \sigma$	Pire des cas : Tchebychev $\mathbb{P}(X - \bar{x} \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \quad k = \frac{1}{\sqrt{\alpha}}$		Cas Gaussien : $\mathbb{P}(X - \bar{x} \geq u_{\alpha/2}\sigma) \leq \alpha$			x	x	X	X
• Fractile et quartile	$\widehat{F}_X(\hat{\phi}_p) = p$	$\widehat{Q}_1 = \hat{\phi}_{\frac{1}{4}}$	$\widehat{Q}_3 = \hat{\phi}_{\frac{3}{4}}$	$DIQ = \widehat{Q}_3 - \widehat{Q}_1$	$MAD = \operatorname{mediane}(x_i - \widehat{M})$				X	X
• Moment et moment centré	$\widehat{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$	$\widehat{m}\widehat{c}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$		$\bar{x} = \widehat{m}_1$ Moyenne $\hat{S}^2 = \widehat{m}\widehat{c}_2$ Variance $\frac{\widehat{m}\widehat{c}_3}{\hat{S}^3}$ Dissymétrie $\frac{\widehat{m}\widehat{c}_4}{\hat{S}^4}$ Aplatissement			x	x	X	X

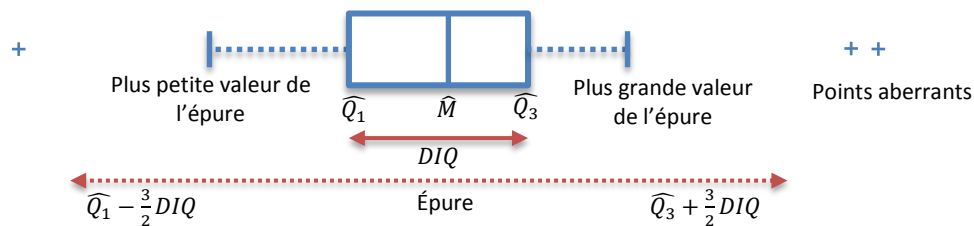
X : peut être utilisé / x : peut être utilisé mais pas toujours très approprié

II. Représentation des variables

1. Camemberts (variables qualitatives)

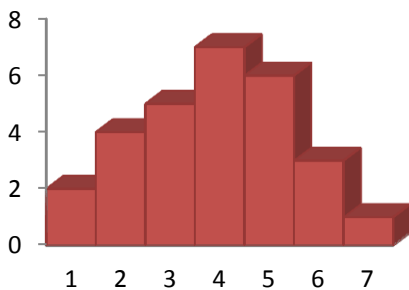


2. Boîtes à moustache (variables quantitatives)



3. Histogramme

Discrétisation d'une variable continue



On doit choisir p bornes b_i . On associe ensuite une hauteur à chaque intervalle $[b_i; b_{i+1}[$.

Choix du nombre p d'intervalles pour n observations :

$$p \geq 1 + \log n$$

règle de Sturges

$$p = \frac{3,5\hat{\sigma}}{\sqrt[3]{n}}$$

règle de Scott

$$p = \frac{2 DIQ}{\sqrt[3]{n}}$$

règle de Freedman Diaconis

Choix des b_i : équirépartition des individus ou des intervalles.