INSA Rouen Méthodes Statistiques M8 **-Formulaire**

- vecteur : $\mathbf{v} =$ est un vecteur colonne de taille n
- transposée $\mathbf{v}^\top = (v_1, \dots, v_n)$ transforme un vecteur colone en un vecteur ligne (et vice versa).
- produit scalaire $p = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i$
- produit extérieur $M = \mathbf{x}\mathbf{y}^\top$ est une matrice
- norme d'un vecteur $\|\mathbf{v}\|^2 = \mathbf{v}^\top \mathbf{v} = \sum_{i=1}^n \ v_i^2$
- matrice M : n lignes et p colonnes
- matrice carré M : $n = p$
- norme matricielle $\|M\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p M_{ij}^2$
- transposée $M^\top$ la matrice telle que $M_{ij}^\top - M_{ji}$
- produit matrice vecteur $\mathbf{u} = M\mathbf{v}$ est un vecteur avec $u_i = \sum_{j=1}^p M_{ij} v_j$
- produit matrice matrice MN
- fonction indicatrice $\mathbb{1}_{\{\Omega\}}$
- probabilité $\mathbb{P}$

- fréquences f_i
- fonction de répartition empirique : $\widehat{F}(x) = \dfrac{1}{n} \sum_{i=1}^n \ \mathbb{1}_{\{x < x_i\}}$
- moyenne : $\overline{x} = \dfrac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \hat{f}_i x_i$
- espérance : $\mathbb{E}(X) = \displaystyle\int x \mathbb{P}(x) dx.$
- variance : $\widehat{\sigma}_x^2 = \dfrac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2$
- médiane : $\mathbb{P}(X < M) = 0,5$
- mode : $\underset{x \in \Omega}{Argmax}\{\mathbb{P}(x)\}$
- fractiles à l'ordre p, $\forall p \in [0,1], \widehat{\Phi}_p$ telle que $\widehat{\mathbb{P}}(X \leq \widehat{\Phi}_p) = p$
- les quartiles :
  - $\widehat{\Phi}_{\frac{1}{4}} = \widehat{Q}_1,$ telle que $\widehat{F}(\widehat{Q}_1) = \frac{1}{4}$,
  - $\widehat{\Phi}_{\frac{1}{2}} = \widehat{Q}_2 = \widehat{M},$ telle que $\widehat{F}(\widehat{M}) = \frac{1}{2}$,
  - $\widehat{\Phi}_{\frac{3}{4}} = \widehat{Q}_3,$ telle que $\widehat{F}(\widehat{Q}_2) = \frac{3}{4}$.
- Distance inter quartile (DIQ) $DIQ = \widehat{Q}_3 - \widehat{Q}_1$
- Moment et moments centrés : $\widehat{m_k} = \dfrac{1}{n} \sum_{i=1}^n x_i^k, \qquad \widehat{mc_k} = \dfrac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$
- épure : $[Q_1 - \frac{3}{2} DIQ, Q_3 + \frac{3}{2} DIQ]$
- combien de classes pour un histogramme ? :
  - règle de Sturges : $p \geq 1 + \log n$
  - règle de Scott : $p \geq \frac{3,5\widehat{\sigma}}{n^{1/3}}$
  - règle de Freedman Diaconis $p \geq 2\frac{DIQ}{n^{1/3}}$
- covariance : $c_{xy} = \dfrac{1}{n} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$
- correlation : $\text{cor}(x, y) = \dfrac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}}$
- probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbf{P}(X = x_i, Y = y_j)}{\mathbf{P}(Y = y_j)}$
- espérance conditionnelle : $\mathbb{E}[Y | X = a] = \displaystyle\sum_{i=1}^n y_i \mathbb{P}(Y = y_i | X = a)$

L'analyse en composantes principales :

- λ valeur propre de la matrice carrée M et $\mathbf{z}$ vecteur propre : $M\mathbf{z} = \lambda\,\mathbf{z}$
- $\|X - \mathbf{u}\mathbf{v}^\top\|_F^2 = \|X\|_F^2 - 2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2$
- $\min\limits_{\mathbf{u},\mathbf{v}} \|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \Leftrightarrow \left\{ \begin{array}{l} \nabla_{\mathbf{u}}\mathcal{J}(\mathbf{u}) = 0 \\ \nabla_{\mathbf{v}}\mathcal{J}(\mathbf{v}) = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -2X\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} = 0 \\ -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} = 0 \end{array} \right. \Leftrightarrow X^\top X\mathbf{v} = \underbrace{\|\mathbf{v}\|^2\|\mathbf{u}\|^2}_{\lambda}\,\mathbf{v}$
- à l'optimum $\|X - \mathbf{u}\mathbf{v}^\top\|_F^2 = \|X\|_F^2 - \lambda$
- axe factoriel $\mathbf{v}$: $X^\top X\mathbf{v} = \lambda\,\mathbf{v}$ et $\|\mathbf{v}\| = 1$
- représentation des individus (composante principale) : $\mathbf{u} = X\mathbf{v}$
- représentation des variables : $\mathrm{cor}(\mathbf{u}, X) = \dfrac{\sqrt{\lambda}}{\sqrt{n}}\,\mathbf{v}$

La régression linéaire :

- modèle linéaire : $y = \sum\limits_{j=1}^{p} x_j\alpha_j + \alpha_0 + \varepsilon$
- estimateur des moindres carrés : $\widehat{\alpha} = (X^\top X)^{-1}X^\top y$