INSA Rouen Méthodes Statistiques M8 **-Formulaire**

Calcul matriciel et notations :
- vecteur : $\mathbf{v}$ = est un vecteur colonne de taille n
- transposée $\mathbf{v}^\top = (v_1, \ldots, v_n)$ 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 $C = AB$, $C_{ij} = \sum_{k=1}^p A_{ik} B_{kj}$
- gradient d'une forme linéaire : $\nabla_{\mathbf{x}} \left(\mathbf{a}^\top \mathbf{x}\right) = \mathbf{a}$
- gradient d'une forme quadratique : $\nabla_{\mathbf{x}} \left(\mathbf{x}^\top A \mathbf{x}\right) = (A + A^\top)\mathbf{x}$
- produit matrice matrice MN
- fonction indicatrice $\mathbb{1}_{\{\Omega\}}$
- probabilité $\mathbb{P}$

Statistiques descriptives :
- fréquences $\widehat{f}_i = \frac{n_i}{n}$ où n est le nombre total d'observations et n_i le nombre d'observant de la modalité i
- fréquences cumulées : $\widehat{F}(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^n \ \mathbb{1}_{\{x < x_i\}}$
- fonction de répartition de la variable aléatoire X : $F(x) = \mathbb{P}(X \leq x)$
- moyenne : $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^n x_i = \sum_{i=1}^n \hat{f}_i x_i$
- espérance : $\mathbb{E}(X) = \displaystyle\int x \mathbb{P}(x) dx.$
- variance : $\widehat{\sigma}_x^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^n (x_i - \overline{x})^2$
- 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 : $\widehat{m_k} = \dfrac{1}{n} \displaystyle\sum_{i=1}^n x_i^k, \qquad \widehat{mc_k} = \dfrac{1}{n} \displaystyle\sum_{i=1}^n (x_i - \bar{x})^k$
- 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 : $c_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

— correlation : $\mathrm{cor}(x, y) = \dfrac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}}$

— 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 : $M\mathbf{z} = \lambda\,\mathbf{z}$

— $\|X\|_F^2 \ = \ \sum_{i=1}^{n}\sum_{j=1}^{p} X_{ij}^2$ Norme de Frobénius de la matrice X

— $\|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \ = \|X\|_F^2 - 2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\ \|\mathbf{v}\|^2$

— $\displaystyle\min_{\mathbf{u},\mathbf{v}} \ \|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \Leftrightarrow \begin{cases} \nabla_{\mathbf{u}}\mathcal{J}(\mathbf{u}) = 0 \\ \nabla_{\mathbf{v}}\mathcal{J}(\mathbf{v}) = 0 \end{cases} \Leftrightarrow \begin{cases} -2X\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} = 0 \\ -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} = 0 \end{cases} \Leftrightarrow \ X^\top X\mathbf{v} = \underbrace{\|\mathbf{v}\|^2\|\mathbf{u}\|^2}_{\lambda}\ \mathbf{v}$

— à l'optimum $\|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \ = \|X\|_F^2 - \lambda$

— axe factoriel $\mathbf{v}$: $X^\top X\mathbf{v} = \lambda\,\mathbf{v} \qquad$ 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 :

1. modèle linéaire : $y = \displaystyle\sum_{j=1}^{p} x_j\alpha_j + \alpha_0 + \varepsilon = X\alpha + \varepsilon; \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
2. estimateur des moindres carrés : $\widehat{\alpha} = (X^\top X)^{-1}X^\top y$
3. estimateur des résidus : $\widehat{e}_i = y_i - \widehat{y}_i \qquad$ avec $\widehat{y} = X\widehat{\alpha}$
4. estimateur de la variance des résidus : $\widehat{\sigma^2} = s^2 = \frac{1}{n-p}\widehat{e}^\top\widehat{e}$
5. $R^2 = \dfrac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$
6. la matrice d'influence : $H = X\ (X^\top X)^{-1}X^\top$
7. les contributions (ou distance de Cook) : $c_i = \frac{H_{ii}}{p(1-H_{ii})^2}\ \frac{\widehat{e}_i^2}{s^2}$

Variables aléatoires et lois

— Soit $Y \ \sim \ \mathcal{N}(0, 1)$ une variable aléatoire normale centrée réduite.
— Soit $Y_1, Y_2, ..., Y_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire.
— **La loi du** χ^2 On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire $Z_n = \sum_{i=1}^{n} Y_i^2$
— **La loi de student** On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad \begin{array}{l} N \sim \mathcal{N}(0, 1) \\ X_n \sim \chi_n^2 \end{array}$$

— **Theorem 0.1 (Théorème du χ^2 (Pearson))** *pour N_{ij} effectif observés et pour n_{ij} effectif théorique*

$$X_{ij} = \frac{N_{ij} - n\ \widehat{\mathbb{P}}_{ij}}{\sqrt{n\ \widehat{\mathbb{P}}_{ij}}} \qquad \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2 \longrightarrow \chi^2_{(I-1)(J-1)}$$

— la variable $T_{n_x+n_y-2}$ suit une loi de student à $n_x + n_y - 2$ degrés de liberté :

$$T_{n_x+n_y-2} = \sqrt{n_x + n_y - 2}\ \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) S_{xy}^2}}$$

avec $S_{xy}^2 = S_x^2 + S_y^2 = \sum_{i=1}^{n_x}(X_i - \overline{X})^2 + \sum_{i=1}^{n_y}(Y_i - \overline{Y})^2$

Mise en œuvre du test du χ^2

1. on construit un tableau de contingence O des observations (2 variables qualitatives de respectivement I et J modalités)
2. on calcule les marginales $p_i = \frac{1}{n}\sum_{j=1}^{J} O_{ij}$
3. on calcule pour chaque case du tableau des effectifs théoriques $T_{ij} = np_ip_j$ (en supposant l'indépendance)
4. on calcule la distance du χ^2 $D(O,T) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - T_{ij})^2}{T_{ij}}$
5. on calcule le nombre de degrés de liberté du χ^2 : $d = (I-1)(J-1)$
6. on regarde dans les tables d'une variable aléatoire Z distribué suivant une loi χ^2 à d degrés de liberté la p-valeur de $D(O,T)$: $\text{pval} = \mathbb{P}\big(Z \geq D(O,T)\big)$
7. on décide qu'on ne peut pas conclure à la dépendance si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Mise en œuvre du test de comparaison des moyennes (T test ou test de student)

1. la question : les deux groupes sont ils des réalisations de la même loi
2. le modèle : gaussien
3. les hypothèses : même variance σ^2 inconnue
4. caclul de

$$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}}$$

$\overline{x}_t$ moyenne avec traitement
$\overline{x}_p$ moyenne sans traitement
$\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\big(\sum_{i=1}^{n_t}(x_{ti} - \overline{x}_t)^2 + \sum_{i=1}^{n_p}(x_{pi} - \overline{x}_p)^2\big)$
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement

5. calcul de la p-valeur $T \sim \mathcal{T}_{n_t+n_p-2}$ (ou lecture sur les tables) $pval = \mathbb{P}(T \leq t)$ ou $pval = \mathbb{P}(T \geq t)$ ou $pval = \mathbb{P}(-|t| \leq T \leq |t|)$
6. on décide que les deux groupes sont ils des réalisation de la même loi si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Mise en œuvre du test sur la pente de la régression

1. les hypothèses : $\begin{cases} \langle_0 : \text{indépendance} & a = 0 \\ \langle_1 : \text{dépendance} & a \neq 0 \end{cases}$
2. caclul de $\widehat{a} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
3. calcul de $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\big(y_i - (\widehat{a}x_i + \widehat{b})\big)^2$ et de $S_x^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$

4. caclul de $\quad t = \frac{\widehat{a}}{\sqrt{\frac{\widehat{\sigma}^2}{S_x^2}}}$
5. calcul du nombre de degrés de liberté $\quad d = n - 2$
6. calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)

$$pval = \mathbb{P}(|T| \geq t)$$

7. on décide de garder $\langle_0$ $(a = 0)$ si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$