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

Définition	Formule	Qualit. binaire	Qualit. multimod	Qualit. ordinale	Quantit. discrète	Quantit. continue
• **Fonction de répartition empirique $\widehat{F_X}$ (cas continu)**	$\widehat{F_X}(x) = \widehat{F_i^c} + (\widehat{F_i^c} - \widehat{F_{i-1}^c})\frac{x - x_{i-1}}{x_i - x_{i-1}}$ $\widehat{F_X}(x_i) = \widehat{F_i^c} = \hat{F}_i - \frac{1}{2}(\hat{F}_i - \hat{F}_{i-1})$					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 :* $\mathbb{E}(X) = \lim_{n\to\infty} \bar{x}$		x	x	X	X
• **Médiane *M***	$\min_{M\in\mathbb{R}} \sum_{i=1}^{n} \lvert x_i - M\rvert$	$\hat{F}_X(M) = \frac{1}{2}$	$\mathbb{P}(X \le M) = \frac{1}{2}$	x			X	X
• **Variance *S***	$\hat{S}^2 = \frac{1}{n}\sum_{i=n}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$	$\hat{s} = \sqrt{\hat{S}^2}$ Ecart type	$S^2 = \mathbb{E}(X - \mathbb{E}^2(X))$		x	x	X	X
• **Fractile et quartile**	$\widehat{F_X}(\hat{\phi}_p) = p$	$\widehat{Q_1} = \hat{\phi}_{\frac{1}{4}}$ $\widehat{Q_3} = \hat{\phi}_{\frac{3}{4}}$	$DIQ = \widehat{Q_3} - \widehat{Q_1}$	$MAD = mediane(\lvert x_i - \widehat{M}\rvert)$			X	X
• **Moment et moment centré**	$\widehat{m_k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$	$\widehat{mc_k} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$	$\bar{x} = \widehat{m_1}$ Moyenne $\hat{S}^2 = \widehat{mc_2}$ Variance $\frac{\widehat{mc_3}}{\hat{S}^3}$ Dissymétrie $\frac{\widehat{mc_4}}{\hat{S}^4}$ Aplatissement		x	x	X	X
• **Dép. :** **(X, Y quant.)**	$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})$ Covariance	$s_X^2 = s_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $s_Y^2 = s_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ Variance	$r = \frac{s_{XY}}{s_X s_Y}$ Corrélation $\in [-1;1]$ $r \approx 0 \Rightarrow x$ indep y $r \approx 1 \Rightarrow x \nearrow y \nearrow$ $r \approx -1 \Rightarrow x \nearrow y \searrow$					
• **Dép. :** **(X qual, Y quant)** **Coef. de détermination**	$S^2_{Y/X} = \frac{\sum_{j\in\Omega_X} n_j(\bar{y}_j - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$							

II. L'ACP

Problème : $\min_{u,v} J(u,v) = \min_{u,v} \|X - uv^{\mathrm{T}}\|_F^2 = \min_{u,v} -2(Xv)^{\mathrm{T}}u + \|u\|^2 + \|v\|^2 = \min_{u,v} \sum_{i}^{n}\sum_{j}^{p}(x_{ij} - u_i v_j)^2$

Solution : $\begin{cases} \nabla_u J(u) = -2Xv + 2\|v\|^2 u = 0 \\ \nabla_v J(v) = -2X^{\mathrm{T}}u + 2\|u\|^2 v = 0 \end{cases} \Leftrightarrow \begin{matrix} Xv = \|v\|^2 u \\ -X^{\mathrm{T}}u = \|u\|^2 v \end{matrix} \Leftrightarrow X^{\mathrm{T}}Xv = \underbrace{\|u\|^2\|v\|^2}_{\lambda} v$

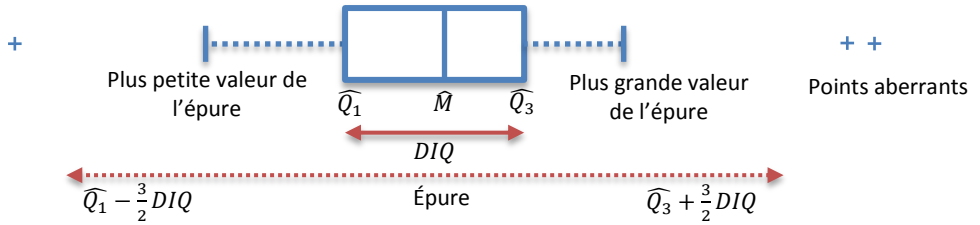
$\Sigma = \frac{X_C{}^{\mathrm{T}} X_C}{n}$ Matrice de covariance

$\rho = \frac{X_R{}^{\mathrm{T}} X_R}{n}$ Matrice de corrélation

$[v, \lambda] = eig(X)$ ACP

$u = Xv$ Données dans l'ACP

$v_n = \sqrt{\frac{\lambda}{n}} v$ Role des variables



Nombre d'intervalles d'un histogramme : $p \ge 1 + \log n$ 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

III. La régression linéaire

1. Le modèle

$$\boxed{\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon}} \quad \boldsymbol{x_i} = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \boldsymbol{\alpha} = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}$$

2. Solution

$$\boxed{\hat{a} = \frac{cov(x,y)}{V_x}} \quad \boxed{\hat{b} = \bar{x}\hat{a} + \bar{y}}$$

Régression simple

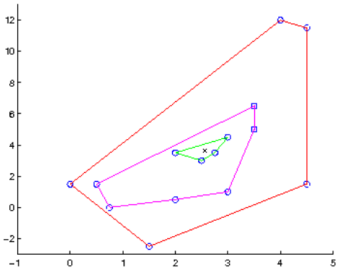
$$X^{\mathrm{T}}\varepsilon = 0 \Rightarrow X^{\mathrm{T}}(y - X\hat{\alpha}) = 0 \Rightarrow (X^{\mathrm{T}}X)\hat{\alpha} = X^{\mathrm{T}}y \Leftrightarrow \boxed{\hat{\alpha} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y}$$

3. Prévision

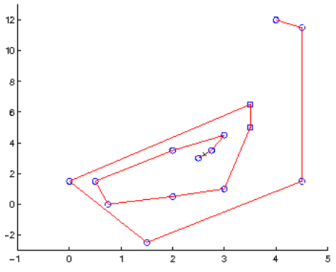
$$z = X\alpha = \underbrace{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{H} y = Hy$$

4. Diagnostic

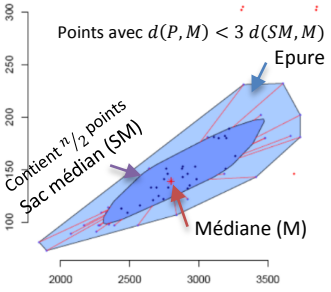
Résultat	Formule		
R²	$\boxed{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SC\ Total} = \underbrace{\sum_{i=1}^{n}(y_i - z_i)^2}_{SC\ Résiduels} + \underbrace{\sum_{i=1}^{n}(z_i - \bar{y})^2}_{SC\ Expliqué}}$	$\boxed{R^2 = \frac{SC\ Expliqué}{SC\ Total}}$ $0 \le R^2 \le 1$ $R = 1$: modèle bon $R = 0$: modèle mauvais	
	$R = \lvert cor(\boldsymbol{y}, \boldsymbol{z})\rvert$ coefficient de corrélation multiple	$R^2 = r_{XY}^2$ en régression simple	
Matrice d'influence	$z = \underbrace{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{H} y = Hy$	$z_i = H_{i\bullet}^T y$	$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ Pour la régression simple
Variances estimées	$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i$	$(s^{(-i)})^2 = \frac{1}{n-p-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i - \frac{\hat{\varepsilon}_i^2}{1 - H_{ii}}\right)$	
Résidus	$\hat{\varepsilon} = y - z$ $\hat{\varepsilon} = (I - H)y$ **Résidus** Sans structure / distrib. normal / pas d'aberrants	$r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1 - H_{ii}}} = \frac{\hat{\varepsilon}_i}{\sqrt{V(\hat{\varepsilon}_i)}}$ **Résidus standardisés**	
Divergence	$\hat{\varepsilon}_i^{(-i)} = \frac{\hat{\varepsilon}_i}{1 - H_{ii}} = y_i - z_i^{-i}$ **Résidus de validation croisée**	$err_{VC} = \sum_{i=1}^{n}\left(\hat{\varepsilon}_i^{(-i)}\right)^2$ **Erreur de validation croisée**	$t_i = \frac{\hat{\varepsilon}_i}{s^{(-i)}\sqrt{1 - H_{ii}}} = \frac{\hat{\varepsilon}_i^{(-i)}}{\sqrt{V\left(\hat{\varepsilon}_i^{(-i)}\right)}}$ **Résidus studentisés** = résidus de VC normalisés
Levier et contribution	$\boxed{levier_i = H_{ii}} = \lVert H_{i\bullet}\rVert^2$ Important si $H_{ii} > \frac{2(p+1)}{n}$ **Levier**	$c_i = \frac{H_{ii}}{p(1 - H_{ii})}\frac{\hat{\varepsilon}_i^2}{s^2}$ Suspect si $c_i > \frac{4}{n}$ **Contribution**	
Cp de Mallows	$\boxed{Cp = \frac{1}{s^2}\sum_{i=1}^{n}\left(y_i - z_i^{(-i)}\right)^2 - n + 2p}$ Conserver la combinaison de variable avec le plus faible Cp		



Médiane de Tukey



Médiane de Jarvis



Sac médian

	2 qualitatives : test du χ^2	1 qualitative, 1 quantitative : test de Student		2 quantitatives : test de Student
Données	$\begin{array}{c\|ccccc} X \backslash Y & \boldsymbol{b_1} & \dots & \boldsymbol{b_J} & \textbf{marge} \\ \hline \boldsymbol{a_1} & n_{11} & \dots & n_{1J} & \boldsymbol{n_{1\bullet}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{a_I} & n_{I1} & \dots & n_{IJ} & \boldsymbol{n_{I\bullet}} \\ \hline \textbf{marge} & \boldsymbol{n_{\bullet 1}} & \dots & \boldsymbol{n_{\bullet J}} & \boldsymbol{n_{\bullet\bullet} = n} \end{array}$	$\begin{array}{cc} \textbf{Y} & \textbf{X} \\ \hline a & x_{a_1} \\ \vdots & \vdots \\ a & x_{a_{n_a}} \\ b & = x_{b_1} \\ \vdots & \vdots \\ b & x_{b_{n_b}} \end{array}$		$\begin{array}{cc} \textbf{Y} & \textbf{X} \\ \hline y_1 & x_1 \\ \vdots & \vdots \\ y_n & x_n \end{array}$ $y_i = ax_i + b + \varepsilon_i$
Hypothèses	$\begin{cases} \mathcal{H}_0 : \text{référence} & \Leftrightarrow \text{ indépendance} \\ \mathcal{H}_1 : \text{alternative} & \Leftrightarrow \text{ dépendance} \end{cases}$	$\begin{cases} \mathcal{H}_0 : \mu_a = \mu_b & X \text{ et } Y \text{ indépendants} \\ \mathcal{H}_{1_1} : \mu_a < \mu_b & \\ \mathcal{H}_{1_2} : \mu_a > \mu_b & X \text{ et } Y \text{ dépendants} \\ \mathcal{H}_{1_3} : \mu_a \neq \mu_b & \end{cases}$		$\begin{cases} \mathcal{H}_0 : \text{indépendance} & \Leftrightarrow a = 0 \\ \mathcal{H}_1 : \text{dépendance} & \Leftrightarrow a \neq 0 \end{cases}$
Modèle	$\mathbb{P}\left((X = x_i) \cap (Y = y_j)\right) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j)$	$\overline{X_a} \sim \mathcal{N}\left(\mu_a, \frac{\sigma^2}{n_a}\right)$ $\overline{X_b} \sim \mathcal{N}\left(\mu_b, \frac{\sigma^2}{n_b}\right)$ $\sigma = \sigma_a = \sigma_b$ même variance		$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n} \widehat{\varepsilon_i}^2 \sim \chi^2_{n-2}$
Statistique	$D_{\chi^2} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \sim \chi^2_{(I-1)(J-1)}$ (eff obs / eff th) $D_{\chi^2} = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\widehat{\mathbb{P}_{ij}} - \widehat{\mathbb{P}_i}\widehat{\mathbb{P}_j})^2}{\widehat{\mathbb{P}_i}\widehat{\mathbb{P}_j}}$	$U = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{\sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \sim \mathcal{N}(0,1)$ Variance connue σ^2	$T = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \sim \mathcal{T}_{n_a+n_b-2}$ Variance inconnue estimée $\hat{\sigma}^2$ $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_a}\left(x_{a_i} - \overline{x_a}\right)^2 + \sum_{i=1}^{n_b}\left(x_{b_i} - \overline{x_b}\right)^2}{n_a + n_b - 2}$	$U = \frac{\hat{a} - a}{\sqrt{\frac{\sigma^2}{s_X^2}}} \sim \mathcal{N}(0,1)$ Variance connue σ^2	$T = \frac{\hat{a} - a}{\sqrt{\frac{\hat{\sigma}^2}{s_X^2}}} \sim \mathcal{T}_{n-2}$ Variance estimée $\hat{\sigma}^2$ $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\widehat{\varepsilon_i}^2}{n-2} = \frac{\sum_{i=1}^{n}\left(y_i - \left(\hat{a}x_i + \hat{b}\right)\right)^2}{n-2}$
Valeur	Calcul de $D_{\chi^2\,obs}$, u, ou t ; valeur de D_{χ^2}, U, ou T à partir des données.			
p-valeur	$p\text{-}val = \mathbb{P}\left(D_{\chi^2_n} \geq D_{\chi^2\,obs}\right)$	$p\text{–}val = \begin{cases} \mathbb{P}(U \leq u) & \text{si } \mathcal{H}_1 = \mathcal{H}_{1_1} \\ \mathbb{P}(U \geq u) & \text{si } \mathcal{H}_1 = \mathcal{H}_{1_2} \\ \mathbb{P}(U \leq -\lvert u\rvert) + \mathbb{P}(U \geq \lvert u\rvert) & \text{si } \mathcal{H}_1 = \mathcal{H}_{1_3} \end{cases}$		$p\text{-}val = \mathbb{P}(\lvert U\rvert \geq u) = \mathbb{P}(U \leq -\lvert u\rvert) + \mathbb{P}(U \geq \lvert u\rvert)$
Décision	**p-valeur :** probabilité d'obtenir un tableau encore plus « rare » au hasard. Se lit généralement dans la table de la loi utilisée. • $p\text{-}val < 5\%$: on garde $\mathcal{H}_1$ *(« peu de tableaux sont plus rares, ce n'est pas du hasard »)* • $p\text{-}val \geq 5\%$: on garde $\mathcal{H}_0$ *(« le tableau n'est pas si rare, ça peut être le hasard »)*			

Loi du χ^2	Thm de Pearson	Loi de Student
$Z_n = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$ $Z_n \sim \chi^2_n$ $Y_i \sim \mathcal{N}(0,1)$ $X_i \sim \mathcal{N}(\mu, \sigma^2)$ $\mathbb{E}(Z_n) = n$ $V(Z_n) = 2n$ $\frac{Z_n - n}{\sqrt{2n}} \to \mathcal{N}(0,1)$	$X_i = \frac{N_i - n\widehat{p_i}}{\sqrt{n\widehat{p_i}}}$ $\sum_{i=1}^{I} X_i^2 \to \chi^2_{I-1}$ $X_{ij} = \frac{N_{ij} - n\widehat{p_{ij}}}{\sqrt{n\widehat{p_{ij}}}}$ $\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2 \to \chi^2_{(I-1)(J-1)}$	$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}}$ $T_n \sim \mathcal{T}_n$ $N \sim \mathcal{N}(0,1)$ $X_n \sim \chi^2_n$ $T_n \to \mathcal{N}(0,1)$

```
%%% Manipulation d'objets %%%

vl = linspace(a, b, n)      % génère n points entre a et b
ma = ones(n,p)   % matrice n x p de 1
n = length(x)    % plus grande dimension de x
[n p] = size(x)  % hauteur et largeur de x

vc = x(:, p)     % pième colonne de la matrice
vl = x(n, :)     % nième ligne de la matrice
ma = x(3:4, 2:5) % sous-matrice (2 x 3)
vc = x(1:2:end, 1);  % matrice contenant un élément sur 2 de la 1ère
colonne de x

vc = diag(x)     % matrice des éléments sur la diagonale de x
vc = find(x > a & x <= b)   % indices des éléments de x correspondants

%%% calculs courants %%%

n = mean(x)   % moyenne de x
n = median(x) % mediana de x
n = mode(x)   % mode de x
n = var(x,1)  % variance de x
n = std(x,1)  % écart type de x
n = min(x)    % minimum de x
n = max(x)    % maximum de x
vc = sum(x)   % somme de x
vc = cumsum(x)       % somme cumulée de x

f = x/sum(x)
F = cumsum(f)
Fc = F - 1/2*( F - [0 F(1:n -1)]); % fct. de répartition empirique (cas
continu)

%% matrice centrée réduite
xcr = (x-ones(n,1)*mean(x))./(ones(n,1)*std(x,1))

%% discretiser x en nbin intervales
d=linspace(min(x),max(x),nbin+1);
for i=1:nbin
       H(i)=length(find(x>d(i)&x<=d(i+1)));
end

%% nombre d'observations a partir d'un échantillon
n = length(a);
va=[]; % valeurs
na=[]; % nb occurences
for i=min(x):max(x)
    p=length(find(x==i));
    if p > 0
        va=[va i];
        na=[na p];
    end
end
```

```
%%% Affichage %%%

plot(x, y, 'o')  % affichage d'une courbe (x, y, paramètres)
bar(x, y)        % histogramme discret (x valeurs, y effectifs)
hist(x, n)       % histogramme continu de x en n espaces
boxplot(x)       % boite a moustache de x
ster(x)          % variation des résidus x

hold on/off      % continuer à dessiner par-dessus la figure
figure(n)        % affichage dans la figure n (évite de réécrire dessus)
subplot(n, p, num)   % met le prochain affichage dans la case num du
subplot n×p
x/ylabel(text);  % titre des axes x et y
axis([minx maxx miny maxy]);       % change les axes
title(text);     % titre de la figure

%%% ACP %%%

Xr = (X - e*mean(X)) % matrice centrée
Xn = Xr./(e*std(X,1))       % matrice centrée réduite
cov = 1/n*(Xr)'*(Xr) % matrice de covariance
cor = 1/n*(Xn)'*(Xn) % matrice de corrélation
[v,d] = eig(Xn'*Xn)% l'ACP
u = Xn*v            % projections sur les axes de l'ACP
Vn = v*((d/n).^(1/2))       % role des variables

%%% Régression linéaire %%%

e = ones(n,1)               % vecteur unitaire
a = (X'*X)\(X'*y)           % coefs de la régression
z = X*a                     % prévision
r = y-z                     % résidus
SCM = sum((z-mean(y)).^2)
SCT = sum((y-mean(y)).^2)
R2 = SCM/SCT                % R²
H = X*(X'*X)^(-1)*X'        % matrice des contributions
h = diag(H)                 % contributions des observations
s2_r = (1/(n-p))*r'*r       % variance estimée
r_s = r./((s2_r*(e-h)).^(1/2))     % résidus standardisés
c = h./(p*(e-h).^2).*r.^2/s2_r     % contributions
err_VC = sum((r.^2)./((e-h).^2))   % erreur de validation croisée

%%% Tests %%%

pval = cdf('norm', u, m, s2)       % loi normale : P(N(m, s2) < u)
pval = cdf('chi2',t,n))   % loi du chi2 : P(D_n < t)
pval = cdf('T',t,n))      % loi de student : P(T_n < t)
```