# Formulaire

SM403

# → CHAPITRE 1

- **Moyenne arithmétique** : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- **Variance** : $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ (Formule de Königs-Huygens)

- **Écart-type** : $\sigma_x = \sqrt{\sigma_x^2}$

→ Mesure la dispersion d'un ensemble de données. Plus il est faible, plus les valeurs sont regroupées autour de la moyenne.

Variable centrée → moyenne = 0
réduite → variance (ou écart-type) = 1

- **Covariance** : $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}$

→ Le signe indique si les variables évoluent dans le même sens ou en sens contraire.

- **Coef. de corrélation** : $r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$

→ Si $r_{xy}$ proche de 1 : évoluent dans le même sens, fortement corrélées
proche de -1 : évoluent dans sens contraire, fortement anti-corr.
proche de 0 : pas dépendantes linéairement, non corrélées

⚠ Corrélation ≠> Causalité

- **Droites de régression** :

| $y = ax + b$ | $x = ay + b$ |
|---|---|
| → $a = \frac{\sigma_{xy}}{\sigma_x^2}$ | → $a = \frac{\sigma_{xy}}{\sigma_y^2}$ |
| → $b = \bar{y} - a\bar{x}$ | → $b = \bar{x} - a\bar{y}$ |

- **Matrice des données centrées** : $X^c = \left[ \frac{x_{ij} - \bar{x_j}}{\sqrt{n}} \right]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ ($\bar{x_j}$ : moyenne de $x_j$)

=> **Matrice de variance - covariance** : $\Sigma = {}^tX^c \times X^c$

→ **Coefs diagonaux** : variances des p variables
→ **Hors - diagonale** : covariances des variables en lignes et en colonnes

Exemple : $\begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_z^2 \end{bmatrix}$

- **Matrice des données centrées réduites** : $X^s = \left[ \frac{x_{ij} - \bar{x_j}}{\sigma_j \sqrt{n}} \right]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ ($\sigma_j$ : écart-type de $x_j$)

=> **Matrice de corrélation** : $R = {}^tX^s \times X^s$

→ **Coefs diagonaux** égaux à 1 => trace de R égale à p
→ **Hors - diagonale** : corrélations des variables en ligne et en colonne

Exemple : $\begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix}$

# → CHAPITRE 2

- **Contribution** : $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100$ → Qge = somme des contributions élevées

- **Qualité** : $qlt = \frac{coord^2(ind\ ;\ \alpha)}{\sum coord^2(ind)}$ ← coordonnée² de l'individu selon l'axe choisi $\alpha$ ; ← somme des coordonnées² de l'individu selon tous les axes

- **Démarche de l'ACP** : On étudie un jeu de données ayant n individus (lignes) et p variables (colonnes).

(1) On réalise une ACP centrée - réduite car les unités entre variables sont différentes, les ordres de grandeurs le sont aussi et de même pour les écarts - types (stats élémentaires).

→ On va donc étudier la matrice des données centrées - réduites $X^s$ (chapitre 1)

(2) On veut désormais trouver les corrélations et anti - corrélations fortes entre les variables : une forte corrélation indique une forte redondance d'informations.

→ On étudie alors la matrice de corrélation R (chap. 1)

⚠ Propriétés de R : Matrice carrée d'ordre p symétrique avec une diagonale de 1 ; coefficients entre -1 et 1.

(3) Puis on vient substituer aux variables initiales les données corrélées en diagonalisant R (car diagonalisable car symétrique). On note dans l'ordre décroissant les valeurs propres de R. Après la diagonalisation, on a $U \Lambda U^{-1}$.

→ On utilise ces valeurs propres pour choisir les axes ayant une forte contribution : calcul de la qge.

⚠ Propriétés des valeurs propres : p valeurs propres, $\sum \lambda_i = p$

(4) On utilise la matrice de changement de base U trouvée lors de la diagonalisation pour en déduire la matrice des composantes principales : $F = X^s \times U$

Cette matrice répertorie les coordonnées de la matrice des données centrées - réduites dans la nouvelle base. Pour étudier les données intéressantes de cette matrice, on calcule la qlt de chaque individu sur chaque axe.

→ On sélectionne les données ayant une qlt proche de 1 (signifie que l'individu est bien représenté par l'axe) et des coordonnées relativement extrêmes.

On représente graphiquement les données sélectionnées pour obtenir le graphe des individus.

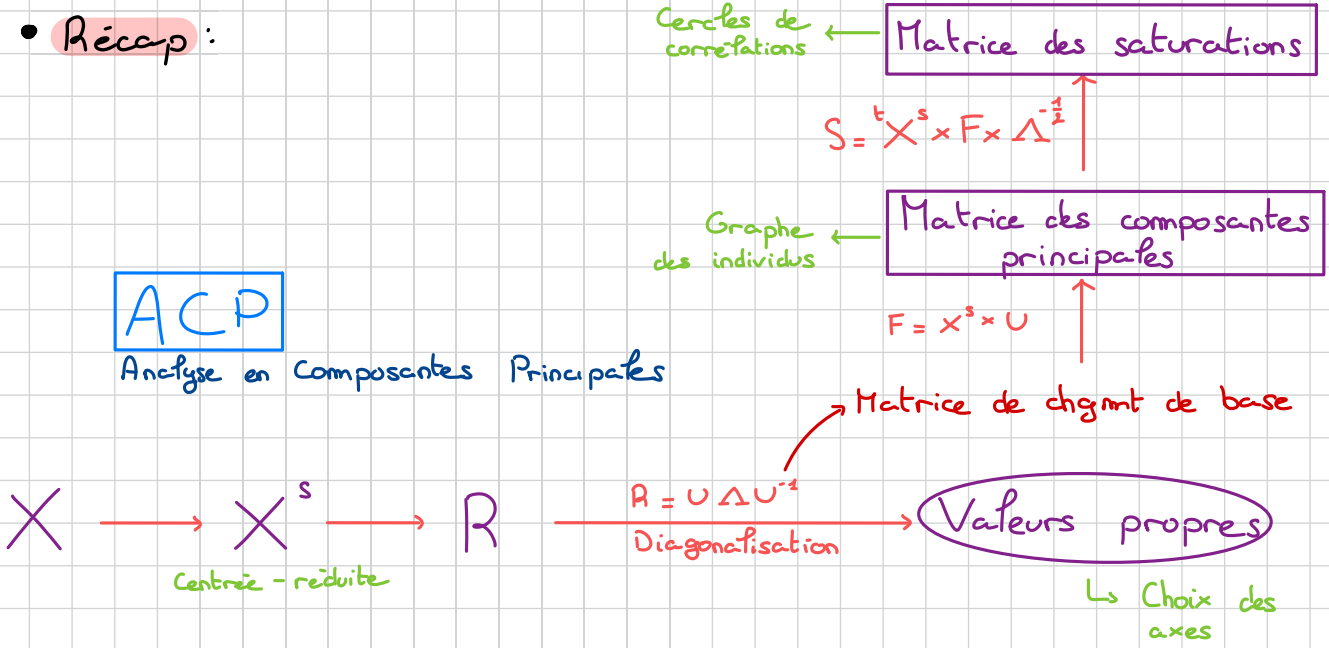
(5) Pour donner un sens aux axes du graphe des individus, on calcule la matrice des saturations : $S = {}^tX^s \times F \times \Lambda^{-\frac{1}{2}}$

Les coefficients sont égaux aux corrélations entre les variables et les axes. On représente graphiquement chaque corrélation dans un cercle de rayon 1 pour obtenir les cercles de corrélation.

→ Plus les variables sont proches du cercle, mieux elles sont représentées dans le plan choisi. Plus elles sont proches d'un axe, mieux elles caractérisent cet axe.

(6) Désormais, on combinant cercle de corrélation et graphe des individus, on en déduit une synthèse de notre jeu de données.

- **Récap :**


Cercles de corrélations
Matrice des saturations
$S = {}^{t}X^{s} \times F \times \Lambda^{-\frac{1}{2}}$
Graphe des individus
Matrice des composantes principales
$F = X^{s} \times U$
Matrice de chgmt de base
ACP
Analyse en Composantes Principales
X
$X^{s}$
Centrée - réduite
R
$R = U \Lambda U^{-1}$
Diagonalisation
Valeurs propres
Choix des axes

# → CHAPITRE 3

- **Échantillonnage**

→ On connait la loi suivie par une population
→ Savoir si un échantillon est représentatif

## → Fréquence empirique

Soit $X \sim \mathcal{B}(p)$, on a $F_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ ← population

Si $n > 30$ ; $np > 5$ et $nq > 5$ ($q = 1-p$) $\Rightarrow F_n \sim \mathcal{N}(p ; \frac{pq}{n})$

Variable standardisée : $F_n^* = \frac{F_n - p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0;1)$

Intervalle de fluctuation : $I_{1-\alpha} = \left[ p - z_\alpha \sqrt{\frac{pq}{n}} \; ; \; p + z_\alpha \sqrt{\frac{pq}{n}} \right]$ ← risque

Rappel : Pour trouver $z_\alpha$

$$P(-z_\alpha \leq X \leq z_\alpha) = 1 - \alpha$$
$$\Leftrightarrow P(X \leq z_\alpha) - P(X \leq -z_\alpha) = 1 - \alpha$$
$$\Leftrightarrow P(X \leq z_\alpha) - (1 - P(X \leq z_\alpha)) = 1 - \alpha$$
$$\Leftrightarrow 2P(X \leq z_\alpha) - 1 = 1 - \alpha$$
$$\Leftrightarrow P(X \leq z_\alpha) = \frac{2-\alpha}{2} \quad \Rightarrow \varphi(z_\alpha) = \frac{2-\alpha}{2}$$

**Valeurs usuelles**

| | |
|---|---|
| $z_{0,10}$ | 1,645 |
| $z_{0,05}$ | 1,96 |
| $z_{0,01}$ | 2,58 |

## → Moyenne empirique

Soit $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$. On a $E(\bar{X}_n) = \mu$ ; $V(\bar{X}_n) = \frac{\sigma^2}{n}$

Si $n > 30$, alors $\bar{X}_n \sim \mathcal{N}(\mu ; \frac{\sigma^2}{n})$

Si $X \sim \mathcal{N}(\mu ; \sigma^2)$, alors $\bar{X}_n \sim \mathcal{N}(\mu ; \frac{\sigma^2}{n})$

Intervalle de fluctuation : $I_{1-\alpha} = \left[ \mu - z_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \mu + z_\alpha \frac{\sigma}{\sqrt{n}} \right]$

→ Variance empirique

Soit $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$. On a $E(S_n^2) = \sigma^2$

Si $X \sim \mathcal{N}(\mu; \sigma^2)$, alors $S_n^2$ suit la loi du $\chi^2$ à $n-1$ ddl

- **Estimation**

→ On connaît les résultats observés sur un échantillon (fréquence, moyenne, variance)
→ Déterminer la meilleure estimation possible en prenant en compte le risque d'erreur

Si continu considérer le "milieu" de chaque intervalle

Estimateurs ponctuels usuels :
- $\hat{\mu} = \bar{x}$
- $\hat{\sigma}^2 = \frac{n}{n-1} \sigma_{obs}^2$

Rappel : $\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$ ; si pondéré $\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$

Intervalles de confiance : Évaluer la confiance que l'on peut avoir en une valeur

- **Pour une proportion** : On utilise $F_n$ pour remplacer p

$$IC_{1-\alpha}(p) = \left[ f - z_\alpha \sqrt{\frac{f(1-f)}{n}} ; f + z_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

- **Pour une moyenne** : On distingue 2 cas

→ $\sigma$ connu : $IC_{1-\alpha}(\mu) = \left[ \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \right]$

→ $\sigma$ inconnu : $IC_{1-\alpha}(\mu) = \left[ \bar{x} - t_{n-1,\alpha} \frac{\sigma_{obs}}{\sqrt{n-1}} ; \bar{x} + t_{n-1,\alpha} \frac{\sigma_{obs}}{\sqrt{n-1}} \right]$

à lire depuis la table de Student $t_\nu$
$p(\{-t_{\nu,\alpha} < t_\nu < +t_{\nu,\alpha}\}) = 1 - \alpha$
→ dernière table en annexe

- **Test d'hypothèse**

→ Confronter une hypothèse à la réalité
→ Réalisation de tests statistiques (bilatéral ou unilatéral)

→ **Test de conformité**

Pour comparer un paramètre de l'échantillon à sa valeur dans la population

- **Moyenne** : On réalise un test bilatéral $\begin{cases} H_0 : \mu_0 = \mu \\ H_1 : \mu_0 \neq \mu \end{cases}$

On distingue 2 cas :

→ **$\sigma^2$ connu** : $Z = \dfrac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0;1)$ sous $H_0$

Région d'acceptation : $[-z_\alpha ; z_\alpha]$

→ **$\sigma^2$ inconnu** : $T = \dfrac{\overline{X_n} - \mu}{\frac{S_n}{\sqrt{n}}} = \dfrac{\overline{X_n} - \mu}{\frac{\sigma_{obs}}{\sqrt{n-1}}}$ suit Student à $n-1$ ddl

Région d'acceptation : $]-t_{n-1,\alpha} ; t_{n-1,\alpha}[$

- **Fréquence** : On réalise le test bilatéral $\begin{cases} H_0 : p_0 = p \\ H_1 : p_0 \neq p \end{cases}$

$Z = \dfrac{F_n - p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0;1)$ sous $H_0$.

Région d'acceptation : $]-z_\alpha ; z_\alpha[$

→ **Test de comparaison**

Pour comparer deux échantillons indépendants de tailles $n_1$ et $n_2$

- **Moyenne** : On réalise le test bilatéral $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$

On distingue les cas suivants :

→ $\sigma_1^2$ et $\sigma_2^2$ connus : $D = \dfrac{\overline{X_{n_1}} - \overline{X_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0;1)$ sous $H_0$

Région d'acceptation : $[-z_\alpha ; z_\alpha]$

→ $\sigma_1^2$ et $\sigma_2^2$ inconnus + inégaux :

- Si $n_1 \geq 30$ et $n_2 \geq 30$ : $D = \dfrac{\overline{X_{n_1}} - \overline{X_{n_2}}}{\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}} \sim \mathcal{N}(0;1)$
- Sinon, on ne connaît pas la loi.

Rappel : $S_1^2 = \dfrac{n_1}{n_1 - 1} \times \sigma_{obs_1}^2$ (de même pour $S_2^2$)

Région d'acceptation : $[-z_\alpha ; z_\alpha]$

→ $\sigma_1^2$ et $\sigma_2^2$ inconnus + égaux : $T = \dfrac{\overline{X_{n_1}} - \overline{X_{n_2}}}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$

où $S^2 = \dfrac{(n_1 - 1) S_{n_1}^2 + (n_2 - 1) S_{n_2}^2}{n_1 + n_2 - 2}$

$T$ suit Student à $n_1 + n_2 - 2$ ddl

Région d'acceptation : $[-t_{n_1+n_2-2;\alpha} ; +t_{n_1+n_2-2;\alpha}]$

- Échantillons appariés : On réalise le test bilatéral $\begin{cases} H_0 : d = 0 \\ H_1 : d \neq 0 \end{cases}$

On a $D$ la variable égale à la différence des deux variables.
On considère $T = \dfrac{\overline{D_n}}{\frac{S_n}{\sqrt{n}}}$ (moyenne empirique) qui suit Student à $n-1$ ddl

Région d'acceptation : $[-t_{n-1;\alpha} ; +t_{n-1;\alpha}]$

- **Fréquences** : On réalise le test bilatéral $\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$

On considère $D = \dfrac{F_{n_1} - F_{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0;1)$

où $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Région d'acceptation : $[-z_\alpha ; z_\alpha]$

---

**Bon à savoir !** Si on cherche dans la table de Student mais que $\nu$ n'est pas indiqué...

Par exemple, cherchons $t_{68;0,05}$ :

| | $\alpha = 0,05$ | | | | | | | | | ? ? | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | 0,126 | 0,255 | 0,388 | 0,529 | 0,681 | 0,851 | 1,050 | 1,303 | 1,684 | 2,021 | 2,423 | 2,704 | 3,551 | ... |
| 80 | 0,126 | 0,254 | 0,387 | 0,527 | 0,679 | 0,848 | 1,046 | 1,296 | 1,671 | 2,000 | 2,390 | 2,660 | 3,460 | |

C'est pas indiqué ?! Pas de panique ! On va estimer la valeur comme suit par proportionnalité de l'augmentation.

Entre 40 et 80 (soit une "distance" de $80-40=40$ depuis $\nu=40$) on a une augmentation de $2,000 - 2,021 = -0,021$.

Comme on cherche pour $\nu = 68$, on cherche l'augmentation entre 40 et 68 (soit une "distance" de $68-40=28$ depuis $\nu=40$).

On procède donc par produit en croix :

| 40 | 28 |
|---|---|
| $-0,021$ | ? → $\dfrac{28\times(-0,021)}{40} = -0,0147$ |

Il ne reste plus qu'à ajouter cette augmentation à la valeur de $t_{40;0,05}$ pour trouver $t_{68;0,05}$ !

$t_{68;0,05} = 2,021 - 0,0147 = \underline{\underline{2,0063}}$