Techniques quantitatives et représentations

BUT - Techniques de Commercialisation
Première année, semestre 2

Hammou El-Otmany

ATER & docteur en mathématiques appliquées
Département de Techniques de commercialisation
Email personnel : hamou.elotmany@gmail.com
Email professionnel : hammou.el-otmany@iut-tarbes.fr
Website : www.hamoelotmany.github.io

Tarbes, 12 Janvier 2022

Déroulement du cours

- 8 séances de 1h30 (cours et TD)
  1. Calcul financier : intérêts simples, intérêts composés, annuités (3h)
  2. Séries à deux caractères : tableaux de contingence et test d'indépendance (2h)
  3. Séries à deux caractères : ajustement linéaire (3h)
  4. Séries chronologiques : saisonnalité, tendance (2h)
- 6 séances de 1h30 en salle informatique (TP)
  1. Mathématiques financières
  2. Séries statistiques bivariées et ajustement linéaire
  3. Séries chronologiques
- Support du cours (accessible sur moodle ou hamoelotmany.github.io)
  1. Présentation contenant 4 chapitres
  2. 4 feuilles de travaux dirigés
  3. Fichier Excel contenant 3 travaux pratiques
- Note du module de mathématiques
  1. 1 note de devoir maison (DM)
  2. 1 note de contrôle continu (CC)
  3. 1 note du projet de statistiques en Excel (PS)

$\leadsto$**Note finale =** $\max$ (PS,moyenne pondérée(CC,DM)) $+$ Bonus "assiduité $\&$ participation"

Descriptif et compétences visées

- **Thématiques :**
  - Calcul financier : intérêts simples, intérêts composés, annuités, ...
  - Statistiques descriptives : séries bivariées, ajustement, corrélation, indépendance des variables, ...
  - Séries chronologiques : données ordonnées dans le temps, tendance, effet saisonnier, ajustement, ...
  - Tableaux de contingence et Test d'indépendance.

- **Compétences visées :** pour les sujets traités dans ce cours, les étudiants sont exceptés pour être en mesure de ce qui suit.
  - Appréhender les incidences des décisions financières
  - Calculer et chiffrer son projet et présenter les indicateurs de suivi et de résultat
  - Évaluer, étudier et simuler les interactions entre caractères
  - Intégrer les comportements liés aux saisons
  - Analyser les statistiques des réseaux sociaux et de sites Internet
  - Mesurer le degré de dépendance entre deux variables

- **Mots clés :** Ajustement linéaire, corrélation, séries chronologiques, saisonnalité, prévisions, liens entre caractères

- **Pré-requis :** Techniques quantitatives et représentations S1

Références pédagogiques utilisées :

- G.C. Rungers et al. *Probabilités et Statistiques Appliquées pour ingénieurs*, 2020, p.378, Lavoisier - technique et documentation
- D. Barnichon. *Mathématiques Et Statistiques Appliquées à l'économie*, 2008
- F., Rosard, J.M., Lagoda. *Statistiques et probabilités appliquées à la gestion et à la finance*, 2016. p.250, Breal.
- R. Veysseyre. *Statistique et probabilités pour les ingénieurs*, 2017, p.400, Aide-mémoire, Dunod.
- C. , Duigou. *Calculs mathématiques, statistiques et financiers avec Excel*, 2016, p. 240. Eni edition.
- J.F., Caulier. *Mathématiques économiques*, 2018, p. 256 De Boeck Sup.

Outline d'exposé

1. **Chapitre 1 : Mathématiques financières**
   - Définitions des intérêts (capitalisation et actualisation)
   - Intérêts simples et intérêts composés
   - Annuités - Tableaux d'amortissement

   $\rightsquigarrow$ Mise en application en TD/TP
2. **Chapitre 2 : Série statistique double ou bivariée**
   - Définition et représentation
   - Tableaux de contingence ou à double entrée
   - Covariance d'une série bivariée
   - Dépendance et indépendance en utilisant la covariance
   - Test d'indépendance χ^2

   $\rightsquigarrow$ Mise en application en TD/TP
3. **Chapitre 3 : Ajustement ou régression linéaire**
   - Définition et représentation d'une série bivariée
   - Ajustement ou régression linéaire
   - Coefficient de corrélation linéaire

   $\rightsquigarrow$ Mise en application en TD/TP
4. **Chapitre 4 : Séries chronologiques**
   - Données ordonnées dans le temps
   - Tendance
   - Effet saisonnière
   - Prévisions

   $\rightsquigarrow$ Mise en application en TD/TP

Chapitre 1 : Mathématiques financières

1. Intérêts, capitalisation et actualisation
2. Annuités - Tableaux d'amortissement
3. Emprunts indivis et emprunts obligatoires

$\rightsquigarrow$ **Mise en application en TD n° 1 et TP n° 1.**

Définition et justification de l'intérêt

Définition

L'intérêt est la rémunération d'un prêt d'argent ou d'un placement versé par période (jour, semaine, quinzaine, mois, année, ...). Autrement, c'est le prix à payer par l'emprunteur au prêteur pour rémunérer le service rendu par la mise disposition d'une somme d'argent pendant une période de temps. Il y en a deux types d'intérêts : Intérêt simple et composé. Pour déterminer le coût d'intérêt, on se réfère aux trois facteurs :

- somme prêtée ou placement capital $VA = C_0$ (valeur actuelle ou présente),
- taux d'intérêts $x\% \in [0; 1]$ par an auquel cette somme est prêtée,
- durée du prêt en t années,

Justification d'intérêt

Plusieurs raisons ont été avancées pour justifier l'existence et l'utilisation de l'intérêt, parmi lesquelles on peut citer :

- La privation de consommation (consommation immédiate, contrepartie)
- La prise en compte du risque (insolvabilité et l'inflation)

Intérêts simples

Définition

Les intérêts simples sont des intérêts qui se calculent uniquement sur le capital initial C_0 pendant la période du prêt ou du placement.

Proposition

Si on réalise un placement C_0 à intérêts simples au taux i pour une période donnée, alors les intérêts produits $I_p(n)$ et le capital cumulé C_f au bout de n périodes sont :

$$I_S(n) = n \times C_0 \times i; \quad C_f = C_0 + n \times C_0 \times i = C_0(1 + n\,i).$$

Le schéma temporel justifie la formule du calcul :

$$\begin{array}{ccccc} 0 & 1 & 2 & \dots & n \\ \downarrow & & & & \downarrow \\ C_0 & & & & C_0 + i\,C_0 + i\,C_0 + \dots i\,C_0 \end{array}$$

Remarques :

- Les intérêts simples sont constants pendant toute la durée du prêt ou du placement.
- Les capitaux successifs augmentent de manière constante pendant la durée du prêt ou du placement (**croissance arithmétique de raison** $r = C_0 \times i$). En effet, ci C_n est le capital à n périodes et C_{n+1} celui obtenu au bout de $(n+1)$ périodes, alors on a :
$$C_{n+1} = C_n + C_0 \times i.$$

Exemples

- Une firme emprunte $12\,000$€ au taux simple de 8% par an. Le capital et les intérêts sont à remboursés dans 3 ans.
  1. Calculer le montant à rembourser à la fin de la 3^{me} année.
  2. Décomposer ce montant en capital et intérêts.
- On place $20\,000$€ pendant 87 jours au taux annuel de 3.9% avec intérêts simples versés quotidiennement.
  1. Calculer les intérêts produits au bout de 87 jours.
  2. Déduire le capital cumulé au bout de 87 jours.

Réponses :

- 1. Data : le capital initial $C_0 = 12\,000$€, le taux périodique (simple) $i = 8\%$ (par an) et la période $n = 3$ (an). Nous avons donc un capital cumulé (valeur future)

$$C_f = C_0(1 + n\,i) = 12\,000\left(1 + 3 \times \frac{8}{100}\right) = 14\,880€.$$

  2. cette valeur acquise en 3 ans, se décompose en $12\,000$€de capital et 2880€d'intérêt.
- 1. Data : le capital initial $C_0 = 20\,000$€, le taux périodique (simple) $i = 3.9\%$ (par an) et la période $n = 87$ (jours.) Les intérêts produits au bout de 87 jours sont

$$I_S(n) = n \times C_0 \times \frac{i}{360} = 87 \times 20\,000 \times \frac{\frac{3.9}{100}}{365}€ \approx 185.92€.$$

  2. Le capital obtenu au bout de 87 jours est $C_f = 20\,000 + 185.920 = 20\,185.92$€.

Exemples

- Une firme emprunte $12\,000$€ au taux simple de 8% par an. Le capital et les intérêts sont à remboursés dans 3 ans.
  1. Calculer le montant à rembourser à la fin de la 3^{me} année.
  2. Décomposer ce montant en capital et intérêts.
- On place $20\,000$€ pendant 87 jours au taux annuel de 3.9% avec intérêts simples versés quotidiennement.
  1. Calculer les intérêts produits au bout de 87 jours.
  2. Déduire le capital cumulé au bout de 87 jours.

Réponses :

- 1. Data : le capital initial $C_0 = 12\,000$€, le taux périodique (simple) $i = 8\%$ (par an) et la période $n = 3$ (an). Nous avons donc un capital cumulé (valeur future)

$$C_f = C_0(1 + n\,i) = 12\,000\left(1 + 3 \times \frac{8}{100}\right) = 14\,880€.$$

  2. cette valeur acquise en 3 ans, se décompose en $12\,000$€de capital et 2880€d'intérêt.
- 1. Data : le capital initial $C_0 = 20\,000$€, le taux périodique (simple) $i = 3.9\%$ (par an) et la période $n = 87$ (jours.) Les intérêts produits au bout de 87 jours sont

$$I_S(n) = n \times C_0 \times \frac{i}{360} = 87 \times 20\,000 \times \frac{\frac{3.9}{100}}{365}€ \approx 185.92€.$$

  2. Le capital obtenu au bout de 87 jours est $C_f = 20\,000 + 185.920 = 20\,185.92$€.

Intérêts composés

Définition

Les intérêts composés sont des intérêts qui se calculent sur le capital initial C_0 et sur les intérêts produits pendant la période du prêt ou du placement.

Proposition

Si on réalise un placement C_0 à intérêts composés au taux i pour une période donnée, alors le capital cumulé C_f après n périodes et les intérêts $I_c(n)$ au bout de n périodes sont :

$$C_f = C_0(1+i)^n; \quad I_C(n) = C_0(1+i)^n - C_0.$$

Le schéma temporel suivant justifie la formule du calcul :

$$\begin{array}{cccccccc}
0 & 1 & 2 & \dots & \dots & \dots & \dots & n \\
\downarrow & \downarrow & \downarrow & & & & & \downarrow \\
C_0 & C_0 + iC_0 & C_1 + i\,C_1 & \dots & \dots & \dots & & C_{n-1} + iC_{n-1} \\
 & C_1 = C_0(1+i) & C_2 = C_0(1+i)^2 & & & & & C_n = C_0(1+i)^n
\end{array}$$

Remarques :

- Les intérêts composés sont variables et augmentent très vite.
- Les capitaux successifs augmentent de manière exponentielle pendant la durée du prêt ou du placement (**croissance géométrique de raison** $q = 1 + i$). En effet, ci C_n est le capital à n périodes et C_{n+1} celui obtenu au bout de $(n+1)$ périodes, alors :
$$C_{n+1} = C_n(1+i) = C_n \times q.$$

Exemples

- Une firme emprunte $12\,000$€ au taux composé de 8% par an. Le capital et les intérêts sont à remboursés après 3 ans.
  1. Calculer le montant à rembourser à la fin de la 3^{me} année.
  2. Décomposer ce montant en capital et intérêts.
  3. Comparer avec l'exemple précédent.
- On place $20\,000$€ pendant 87 jours au taux annuel de 3.9% avec intérêts composés versés quotidiennement.
  1. Calculer le capital cumulé au bout de 87 jours.
  2. Déduire les intérêts produits au bout de 87 jours.

Réponses :

- 1. Data : le capital initial $C_0 = 12\,000$€, le taux périodique (composé) $i = 8\%$ (par an) et la période $n = 3$ (an). Nous avons donc un capital cumulé (valeur future)

$$C_f = C_0(1+i)^n = 12\,000\left(1+\frac{8}{100}\right)^3 \approx 15\,116,54€.$$

  2. cette valeur acquise en 3 ans, se décompose en $12\,000$€de capital et $3116,54$€ d'intérêt.
  3. Comme l'intérêt est composé, le loyer de l'argent est plus élevé que dans le cas d'intérêt simple (En excel, la fonction VC (valeur cumulée) donne directement le résultat souhaité).
- 1. Le capital cumulé au bout de 87 jours est $C_f = 20\,000 \times \left(1+\frac{\frac{3,9}{100}}{360}\right)^{87} \approx 20\,189,38$€.
  2. Les intérêts produits au bout de 87 jours sont

$$I_C(n) = C_f - C_0 = 20\,189,38 - 20\,000 = 189,30€.$$

Exemples

- Une firme emprunte $12\,000$€ au taux composé de 8% par an. Le capital et les intérêts sont à remboursés après 3 ans.
  1. Calculer le montant à rembourser à la fin de la 3^{me} année.
  2. Décomposer ce montant en capital et intérêts.
  3. Comparer avec l'exemple précédent.
- On place $20\,000$€ pendant 87 jours au taux annuel de 3.9% avec intérêts composés versés quotidiennement.
  1. Calculer le capital cumulé au bout de 87 jours.
  2. Déduire les intérêts produits au bout de 87 jours.

Réponses :

- 1. Data : le capital initial $C_0 = 12\,000$€, le taux périodique (composé) $i = 8\%$ (par an) et la période $n = 3$ (an). Nous avons donc un capital cumulé (valeur future)

  $$C_f = C_0(1+i)^n = 12\,000\left(1+\frac{8}{100}\right)^3 \approx 15\,116,54€.$$

  2. cette valeur acquise en 3 ans, se décompose en $12\,000$€de capital et $3116,54$€ d'intérêt.
  3. Comme l'intérêt est composé, le loyer de l'argent est plus élevé que dans le cas d'intérêt simple (En excel, la fonction VC (valeur cumulée) donne directement le résultat souhaité).
- 1. Le capital cumulé au bout de 87 jours est $C_f = 20\,000 \times \left(1+\frac{\frac{3,9}{100}}{360}\right)^{87} \approx 20\,189,38€.$
  2. Les intérêts produits au bout de 87 jours sont

  $$I_C(n) = C_f - C_0 = 20\,189,38 - 20\,000 = 189,30€.$$

Propriétés

- La valeur actuelle (VA) ou présente C_0 d'un capital C_n est : $VA = C_0 = \dfrac{C_n}{(1+i)^n}$.
- Le taux périodique composé est : $i = \left(\dfrac{C_n}{C_0}\right)^{1/n} - 1$
- La durée d'un investissement est : $n = \dfrac{\log\left(\dfrac{C_n}{C_0}\right)}{\log(i+1)}$

Exemples

- Calculer la valeur actuelle C_0 d'un capital placé à intérêts composés aujourd'hui au taux annuel de $1,25\%$ qui aura la valeur acquise de 12000€ dans 12 ans.
- On place 1200€ durant 1 an. On accumule ainsi 180€ d'intérêts. Calculer le taux d'intérêt nominal de ce placement sachant que les intérêts sont composés chaque trimestre.
- On place 1200€ dans un compte fermé rémunéré à un taux composé de 9% par an. Combien de temps faut-il pour doubler ce montant ?

Réponses :

- On applique la formule de la valeur actuelle $VA = \dfrac{C_n}{(1+i)^n}$ avec $n = 12$, $i = 1.25\%$ et $C_n = 12000€$. On obtient $VA = C_0 = \dfrac{12000}{(1+\frac{1.25}{100})^{12}} \approx 10338.10€$. Il faut donc placer aujourd'hui $C_0 \approx 10338.10€$ et patienter 12 ans pour accumuler 12 000. En d'autres termes, on ne touche pas aux fruits du placement pendant 12 ans.
- On applique la formule $i = \left(\dfrac{C_n}{C_0}\right)^{1/n} - 1$ avec $C_0 = 1200€$, $C_4 = 1180€$ et $n = 4$ (trimestres). On obtient : $i = \left(\frac{1200+180}{1200}\right)^{1/4} - 1 \approx 0.0356 = 3.56\%$. par trimestre. Par conséquent, le taux nominal est $j = 4 \times i = 4 \times 3.56\% = 14.24\%$ par an (en Excel, utiliser directement la fonction TAUX).
- On sait que $C_0 = 1200€$, $i = 9\%$ et $C_n = 2 \times 1200 = 2400€$ sauf que la durée n est inconnue. On a $n = \dfrac{\log\left(\dfrac{C_n}{C_0}\right)}{\log(i+1)} = \dfrac{\log\left(\dfrac{2400}{1200}\right)}{\log(\frac{9}{100}+1)} \approx 8.04 \approx 8$ ans.

Remarque : Le taux nominal ou de marché est le taux permettant de définir les intérêts à rembourser en fonction de l'emprunt.
Le taux annuel effectif global (TAEG) est un taux prenant en compte les frais de la souscription de l'emprunt.

Taux proportionnel et taux équivalent

Définitions

- Le taux proportionnel au taux i pour une sous-période est le taux qui est appliqué à intérêts simples sur toutes les sous-périodes composant la période aboutit à la même valeur acquise que celle obtenue en appliquant le taux i sur la période.
- Le taux équivalent au taux i pour une sous-période est le taux qui est appliqué à intérêts composés sur toutes les sous-périodes composant la période aboutit à la même valeur acquise que celle obtenue en appliquant le taux i sur la période.

Méthodes de calcul

- On divise la périodes en k sous-périodes et on veut calculer le taux proportionnel au taux i pour une sous-période. Notons, i_k ce taux proportionnel, alors on a

$$i_k = \frac{i}{k}.$$

- On divise la périodes en k sous-périodes et on veut calculer le taux équivalent au taux i pour une sous-période. Notons i_k ce taux équivalent, alors on a

$$i_k = (1+i)^{1/k} - 1.$$

Exemples

- On place un capital 290€ pendant sept mois à intérêts simples au taux annuel $4,08\%$.
  1. Calculer le taux proportionnel mensuel.
  2. Calculer la somme à rembourser.
- On place un capital 290€ pendant sept mois à intérêts composés au taux annuel $4,08\%$.
  1. Calculer le taux équivalent mensuel.
  2. Calculer la somme à rembourser.

Réponses :

- 1. On a $i = 0,0408$. La période est l'année, la sous-période est le mois donc $k = 12$. Le taux proportionnel mensuel est donc

$$i_{12} = \frac{0,0408}{12} = 0.0034 = 0,34\%$$

  2. La somme à rembourser est $C_7 = 290(1 + 7 \times 0.0034) \approx 296,90$€.
- 1. On a $i = 0,0408$. La période est l'année, la sous-période est le mois donc $k = 12$. Le taux proportionnel mensuel est donc

$$i_{12} = (1 + 0,0408)^{1/12} - 1 = 0,003338 = 0,338\%$$

  2. La somme à rembourser est $C_7 = 290(1 + 0,003338)^7 \approx 296,84$€.

Exemples

- On place un capital $290€$ pendant sept mois à intérêts simples au taux annuel $4,08\%$.
  1. Calculer le taux proportionnel mensuel.
  2. Calculer la somme à rembourser.
- On place un capital $290€$ pendant sept mois à intérêts composés au taux annuel $4,08\%$.
  1. Calculer le taux équivalent mensuel.
  2. Calculer la somme à rembourser.

Réponses :

- 1. On a $i = 0,0408$. La période est l'année, la sous-période est le mois donc $k = 12$. Le taux proportionnel mensuel est donc

$$i_{12} = \frac{0,0408}{12} = 0.0034 = 0,34\%$$

  2. La somme à rembourser est $C_7 = 290(1 + 7 \times 0.0034) \approx 296,90€$.
- 1. On a $i = 0,0408$. La période est l'année, la sous-période est le mois donc $k = 12$. Le taux proportionnel mensuel est donc

$$i_{12} = (1 + 0,0408)^{1/12} - 1 = 0,003338 = 0,338\%$$

  2. La somme à rembourser est $C_7 = 290(1 + 0,003338)^7 \approx 296,84€$.

Emprunts et annuités

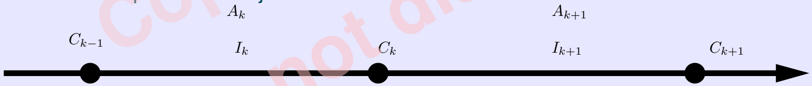
- Les investissements sont financés par des emprunts, qui sont ensuite remboursés par annuités ou mensualités.
- Une annuité est une somme d'argent versée à chaque période, en général d'un an par un épargnant pour constituer une épargne ou par un emprunteur pour rembourser un emprunt ou une dette.
- Un emprunt indivis est un emprunt souscrit auprès d'un et un seul prêteur et faisant l'objet d'un contrat engageant l'emprunteur à verser n annuités au préteur pour rembourser le capital prêté ainsi que les intérêts calculées au taux $i\%$ du contrat.

Propriétés

Supposons que l'on emprunte un capital C au taux annuel composé i, que l'on rembourse au bout de n années par échéance constantes. Alors, le montant de chaque annuité est

$$a = \frac{i \times C_0}{1-(1+i)^{-n}}.$$

Le schéma temporel suivant justifie la formule du calcul :



Remarque : Il existe deux modes de calcul des remboursements : l'annuité constante (la somme d'argent versée à chaque période est la même) et l'amortissement constant (annuité dégressive ou variable, peu utilisée).

Tableau d'amortissement

D'après le schéma temporel, chaque annuité a_k se décompose en :

- un amortissement A_k servant à rembourser une part du capital empruntél
- un paiement des intérêts de la période $I_k = C_{k-1}$ avec C_{k-1} le capital restant dû au début de la période correspondant au k^e versement.

Période n°k	Capital restant dû C_k^{de} en début d'exercice	Intérêts I_k	Capital amorti A_k	Annuité a_k	Capital restant dû C_{k-1} en fin d'exercice
1	$C_0^{de} = C_0$	$I_1 = iC_0$	$A_1 = a_1 - I_1$	$a_1 = A_1+I_1$	$C_1 = C_0 - A_1$
2	$C_2^{de} = C_1$	$I_2 = iC_1$	$A_2 = a_2 - I_2$	$a_2 = A_2+I_2$	$C_2 = C_1 - A_2$
k	$C_k^{de} = C_{k-1}$	$I_k = iC_{k-1}$	$A_k = a_k - I_k$	$a_k = A_k+I_k$	$C_k = C_{k-1} - A_k$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$C_n^{de} = C_{n-1}$	$I_n = iC_{n-1}$	$A_n = a_n - I_n$	$a_n = A_n+I_n$	$C_n = C_{n-1} - A_n$

Remarque :

- On a $C_0 = A_1 + A_2 + \cdots + A_n$.
- Pour tout $k \in \{0, 1, \cdots, n-1\}$, on a $C_k = C_0 - \sum_{i=1}^{k} A_i$ et $C_{n-1} = A_n$.
- Pour tout $k \in \{0, 1, \cdots, n-1\}$,, on a $a_k - a_{k+1} = A_{k+1} - (1+i)A_k$ et $a_n = (1+i)m_n$.
- Annuités constantes $\Longrightarrow A_1 = \frac{i \times C_0}{(1+i)^n - 1}$; $a = a_k = \frac{i \times C_0}{1-(1+i)^{-n}}$ et $A_{k+1} = (1+i)A_k$ pour tout $k \in \{0, 1, \cdots, n-1\}$.

Application numérique

On réalise un emprunt C_0 de 120000€ auprès de la banque sur 4 ans à un taux annuel $2,8\%$. Faire le tableau des annuités de cet emprunt en envisageant l'amortissement variable et l'annuité constante.

- Ici, l'annuité est constante, donc chaque annuité de remboursement s'élèvera à

$$a = a_k = \frac{0.028 \times 120000}{1-(1+0.028)^{-4}} \approx 32128.99\text{€}, \text{ pour tout } k \in \{0,1,2,3,4\}$$

$$I_1 = i \times C_0 = \frac{2.8}{100} \times 120000 = 3360\text{€},\ A_1 = a_1 - I_1 = 32128.99 - 3360 = 28768.99\text{€}.$$

Année k	**Capital restant dû** C_k^{de} **en début d'exercice**	**Intérêts** I_k	**Capital amorti** A_k	**Annuité** a_k	**Capital restant dû** C_k **en fin d'exercice**
1	120000.00	3360.00	28768.99	32128.99	91231.01
2	91231.01	2554.47	29574.52	32128.99	61656.49
3	61656.49	1726.38	30402.61	32128.99	31253.88
4	31253.88	875.11	31253.88	32128.99	0

- Le montant total à rembourser est

$$M_T = \sum_{i=1}^{4} a_k = 4 \times a = 4 \times 32128.99 \approx 128515.96\text{€}.$$

- Le coût de l'emprunt s'élève donc à $128515.96 - 120000 = 8515.96$€ (=somme des intérêts=capital remboursé - capital emprunté).

Chapitre 2 : séries statistiques à deux variables

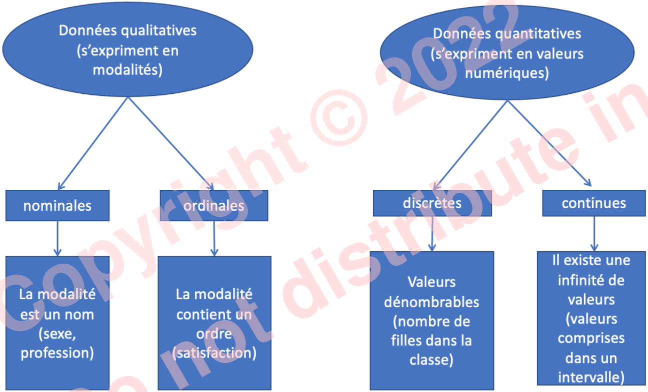
1. Introduction aux séries à deux variables
2. Définition et représentation
3. Mesures d'une série statistique à deux caractères
4. Tableaux de contingence ou à double entrée
5. Test d'indépendance du χ^2

Ce cours est basé sur les notions de séries statistiques à un caractère (univariées) vues en premier semestre

$\rightsquigarrow$ **Mise en application en TD n° 2 et TP n° 2.**

Définition et exemple d'une série statistique à deux variables

- La statistique descriptive à deux variables (à deux caractères ou bivariées) est une série statistique qui caractérise les relations existantes entre deux séries d'observations considérées simultanément où les deux caractères peuvent être de nature qualitative ou quantitative.



Exemple 1.

Durant la première année de Covid-19, un constructeur français a lancé la production d'une voiture que l'on peut équiper de plusieurs options. On note X le nombre de voitures vendues et Y le nombre d'options vendues selon les régions.

Régions	1	2	3	4	5	6
X	1300	3200	6000	2900	3700	4800
Y	550	1400	3200	1350	1715	2500

La série statistique Z composée des couples $(1300; 550)$, $(3200; 1400)$, $(6000; 3200)$,$(2900; 1350)$, $(3700; 1715)$ et $(4800; 2500)$ est une série statistique à deux variables X et Y que notera dans la suite par $Z = (X, Y)$.

Remarques :

- Dans certains cas, il existe une relation entre les deux variables X et Y (ex. consommation et vitesse d'une voiture, poids et taille, chiffre d'affaires et les charges)
- La relation entre deux variables ne signifie pas que l'un est la cause de l'autre. Elle peut être due au hasard ou à l'effet d'une $3^{\text{ème}}$ variable cachée (ex. la vente de lunettes de soleil liée à la vente des cônes glacés sans qu'aucun des deux ne soit la cause ou l'effet de l'autre $\hookrightarrow$ conséquence de la météo).

Mesures d'une série statistique bivariée

Définitions : point moyen d'une série statistique bivariée

Soit (X, Y) une série statistique quantitative observée sur une même population de taille n. Le **point moyen** (qui mesure la **tendance centrale** de (X, Y) est le couple $(\bar{x}, \bar{y})$ où

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (\text{données simples } (x_i, y_i)_{1\leq i\leq n})$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{p} n_i y_i \quad (\text{données groupées } (x_i, y_i, n_i)_{1\leq i\leq p} \text{ où } n = \sum_{i=1}^{p} n_i)$$

Exemple : reprenons l'exemple 1. du constructeur de voitures

Avec les notations ci-dessus, on a

$$\bar{x} = \frac{1300 + 3200 + 6000 + 2900 + 3700 + 4800}{6} = 3650$$

$$\bar{y} = \frac{550 + 1400 + 3200 + 1350 + 1715 + 2500}{6} = 1785.833$$

Ainsi, le point moyen de la série $Z = (X, Y)$ est $(\bar{X}, \bar{Y}) = (3650, 1785.833)$.

Définition : covariance entre X et Y

Soit (X, Y) une série statistique quantitative observée sur une même population de taille n. On appelle **covariance** de (X, Y) la moyenne des produits des écarts aux moyennes et on note $Cov(X, Y)$ le nombre

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \quad (\text{données simples } (x_i, y_i))$$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i y_i - \bar{x}\,\bar{y} \quad (\text{données groupées } (x_i, y_i, n_i)_{1\leq i\leq p})$$

où $\bar{x}$ (resp. $\bar{y}$) est la moyenne de X (resp. de Y).

Remarque : La covariance (généralisation de la variance à deux variables) mesure la dispersion, des valeurs autour du point moyen $(\bar{x}, \bar{y})$.

Exemple : reprenons l'exemple 1. du constructeur de voitures

Avec les notations ci-dessus, on a :

$$\frac{1}{6}\sum_{i=1}^{6} x_i y_i = \frac{(1300 \times 550) + (3200 \times 1400) + (6000 \times 3200) + (2900 \times 1350) + (3700 \times 1715) + (4800 \times 2500)}{6} = 7775916,667$$

donc $Cov(X,Y) = 7775916.667 - 3650 \times 1785.833 \approx 1257625.$

Tableau de contingence - définition et exemple

Définition

- Le **tableau de contingence** (à double entrée) est un tableau de comptage croisant les modalités ou les valeurs de deux caractères qui nous permet de présenter une série statistique bivariée (à deux dimensions).
- Dans un tableau de contingence, il existe plusieurs possibilités de croisement.

Caractère 2 / Caractère 1		Valeurs		Modalités	
		simples	groupées	simples	groupées
Valeurs	simples	1	5	12	27
	groupées	3	10	11	2
Modalité	simples	2	18	15	11
	groupées	3	7	9	20

Table – Forme générique d'un tableau de contingence ou des effectifs conjoints ou à double entrée

Distribution conjointe en effectif et en fréquence

Définition

- La **distribution d'un caractère** est la liste des modalités $(x_1, \cdots, x_k)$ du caractère, chacune étant associée à son effectif ou à sa fréquence dans l'échantillon.
- La **distribution conjointe en effectif** n_{ij} est le nombre de fois où la modalité x_i de X et la modalité y_j de Y ont étés observées simultanément.
- La **distribution conjointe en fréquence** f_{ij} est la proportion $\frac{n_{ij}}{n}$ ou le pourcentage $\frac{n_{ij}}{n} \times 100\%$ des individus possédant à la fois la modalité x_i de X et la modalité y_j de Y.

En général, le tableau de contingence fait apparaître les modalités x_i de la série statistique X et en colonnes les modalités y_j de la série statistique Y.

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_m	*Total*
x_1	n_{11}	$\cdots$	n_{1j} ou f_{1j}	$\cdots$	n_{1m} ou f_{1m}	$n_{1\cdot}$ ou $f_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	$\vdots$	$\ddots$	n_{ij} ou f_{ij}	$\ddots$	$\vdots$	$n_{i\cdot}$ ou $f_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1} ou f_{k1}	$\cdots$	n_{kj} ou f_{kj}	$\cdots$	n_{km} ou f_{km}	$n_{k\cdot}$
Total	$n_{\cdot 1}$ ou $f_{\cdot 1}$	$\cdots$	$n_{\cdot j}$ ou $f_{\cdot j}$	$\cdots$	$n_{\cdot m}$ ou $f_{\cdot m}$	n

Exemple

Lors d'une enquête, on a demandé aux personnes interrogées dans une entreprise leurs nombre d'enfants à charge X et leur revenus mensuels Y qu'ils perçoivent en centaines d'euros.

- Distribution en effectifs :

X \ Y	$[0;8[$	$[8;15[$	$[15;25[$	$[25;30[$	$[35;40[$	$[40;45[$	$[45;50[$	*Total*
1	8	13	14	15	2	0	0	52
2	6	15	10	9	3	0	0	43
3	4	5	10	8	1	0	0	28
4	2	3	9	3	3	3	1	24
5	2	1	5	3	3	2	1	17
Total	22	37	48	38	12	5	2	164

- Distribution en fréquence :

X \ Y	$[0;8[$	$[8;15[$	$[15;25[$	$[25;30[$	$[35;40[$	$[40;45[$	$[45;50[$	*Total*
1	$\frac{8}{164} \approx 0.048$	0.079	0.085	0.091	0.012	0	0	0.315
2	0.036	0.091	0.060	0.054	0.018	0	0	0.259
3	0.024	0.030	0.060	0.048	0.006	0	0	0.168
4	0.012	0.018	0.054	0.018	0.018	0.018	0.006	0.144
5	0.012	0.006	0.030	0.018	0.018	0.012	0.006	0.102
Total	0.134	0.224	0.289	0.229	0.072	0.030	0.012	[illegible]

Distributions marginales - effectifs marginaux

Définition

- La **distribution marginale** est une distribution statistique à un seul caractère donnant la répartition des modalités selon le 1^{er} caractère indépendamment du 2^{nd} caractère.
- L'**effectif marginal** de la modalité x_i de X (resp. de la modalité y_j de Y) est :

$$\forall i \in \{1, \ldots, k\}, \quad n_{i\cdot} = \sum_{j=1}^{m} n_{ij} \quad \left(\text{resp. } \forall j \in \{1, \ldots, m\}, \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}\right).$$

- L'**effectif total** (la taille de l'échantillon) est défini par :

$$n = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} = \sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{m} n_{\cdot j}.$$

X \ Y	$[0;8[$	$[8;15[$	$[15;25[$	$[25;30[$	$[35;40[$	$[40;45[$	$[45;50[$	*Total*
1	8	13	14	15	2	0	0	52
2	6	15	10	9	3	0	0	43
3	4	5	10	8	1	0	0	28
4	2	3	9	3	3	3	1	24
5	2	1	5	3	3	2	1	17
Total	22	37	48	38	12	5	2	164

Distributions marginales - fréquences marginales

Définition

- La **fréquence marginale** est la proportion des individus possédant la modalité x_i (resp. y_j de Y) de X indépendamment de Y (resp. de X) :

$$\forall i \in \{1, \ldots, k\}, \quad f_{i\cdot} = \frac{n_{i\cdot}}{n} = \sum_{j=1}^{m} \frac{n_{ij}}{n} \quad \left(\text{resp. } \forall j \in \{1, \ldots, m\}, \quad f_{\cdot j} = \frac{n_{\cdot j}}{n} = \sum_{i=1}^{k} \frac{n_{ij}}{n}\right).$$

X \ Y	[0; 8[	[8; 15[	[15; 25[	[25; 30[	[35; 40[	[40; 45[	[45; 50[	*Total*
1	0.048	0.079	0.085	0.091	0.012	0	0	0.315
2	0.036	0.091	0.060	0.054	0.018	0	0	0.259
3	0.024	0.030	0.060	0.048	0.006	0	0	0.168
4	0.012	0.018	0.054	0.018	0.018	0.018	0.006	0.144
5	0.012	0.006	0.030	0.018	0.018	0.012	0.006	0.102
Total	0.134	0.224	0.289	0.229	0.072	0.030	0.012	1

Distributions marginales

- Distributions marginales de X en effectifs (enfants à charge)

X	
1	52
2	43
3	28
4	24
5	17
Total	164

- Distribution marginales de X en fréquence (enfants à charge)

X	
1	0.315
2	0.259
3	0.168
4	0.144
5	0.102
Total	1

- Distribution marginale de Y en effectifs (revenus mensuels)

Y	
$[0; 8[$	22
$[8; 15[$	37
$[15; 25[$	48
$[25; 30[$	38
$[35; 40[$	12
$[40; 45[$	5
$[45; 50[$	2
Total	164

- Distribution marginales de Y en fréquence (revenus mensuels)

Y	
$[0; 8[$	0.134
$[8; 15[$	0.224
$[15; 25[$	0.289
$[25; 30[$	0.229
$[35; 40[$	0.072
$[40; 45[$	0.030
$[45; 50[$	0.012
Total	1

Distributions marginales - moyenne marginale et variance marginale

Définition

- La **moyenne marginale** de X (resp. de Y) est définie par :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i}.x_i \quad \left(\text{resp. } \bar{Y} = \frac{1}{n}\sum_{j=1}^{m} n_{.j}y_j\right)$$

- La **variance marginale** de X (resp. de Y) est définie par :

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i}.(x_i - \bar{X})^2 \left(\text{resp. } S_Y^2 = \frac{1}{n}\sum_{j=1}^{m} n_{.j}(y_j - \bar{Y})^2\right)$$

Exemple

Reprenons notre situation de référence. Par définition, on a

$$\bar{X} = \frac{52 \times 1 + 43 \times 2 + 28 \times 3 + 24 \times 4 + 17 \times 5}{164} =$$

$$\bar{Y} = \frac{22 \times 4 + 37 \times 11.5 + 48 \times 20 + 38 \times 27.5 + 12 \times 37.5 + 5 \times 42.5 + 2 \times 47.5}{164} =$$

Distributions conditionnelles

Définition

- La **distribution conditionnelle** est une distribution statistique lorsque l'on a restreint la série statistique étudiée à une seule modalité (distribution univariée).
- On appelle distribution conditionnelle de X (resp. de Y) sachant $Y = y_j$ (resp. $X = x_i$) la distribution :

$$\{(x_1, n_{1j}), \cdots, (x_k, n_{kj})\} \quad (\text{resp. } \{(y_1, n_{i1}), \cdots, (y_m, n_{im})\}) \tag{1}$$

$\hookrightarrow$ Il s'agit donc de la distribution à une variable donnée par la i-ème ligne (resp. la j-ème colonne) du tableau de contingence.

- Distribution conditionnelle de X sachant $Y = [8; 15[$

$X_{\|Y=[8;15[}$	
1	13
2	15
3	5
4	3
5	1
Total	37

- Distribution conditionnelle de Y sachant que $X = 3$

$Y_{\|X=3}$	
[0; 8[	4
[8; 15[	5
[15; 25[	10
[25; 30[	8
[35; 40[	1
[40; 45[	0
[45; 50[	0
Total	18

Dépendance entre deux variables

Définition

Soit (X,Y) une série statistique quantitative observée sur une même population de taille n dont les fréquences du i^{eme} individus sont respectivement $f_{i\cdot}$ et $f_{\cdot j}$. Pour $1 \leq i,j \leq n$, on note f_{ij} la fréquence du couple (x_i, y_j) (voir le tableau de contingence).

- On dit que les variables X et Y sont **indépendantes** si pour tout $1 \leq i,j \leq n$,

$$f_{ij} = f_{i\cdot} \times f_{\cdot j}$$

- On dit que les variables X et Y sont **dépendantes** si elles ne sont pas indépendantes, c'est-à-dire il existe $1 \leq i,j \leq n \ : \ f_{ij} \neq f_{i\cdot} \times f_{\cdot j}$.

Propriétés

Soit (X,Y) une série statistique quantitative observée sur une même population de taille n. Si X et Y sont indépendantes, alors $Cov(X,Y) = 0$ et $V(X+Y) = V(X) + V(Y)$.

Remarques :

- $Cov(X,Y)$ n'implique pas nécessairement que X et Y sont indépendantes.
- $E(X+Y) = E(X) + E(Y)$ est **toujours** vraie (que X et Y soient indépendantes ou pas).
- Si X et Y sont dépendantes alors $V(X+Y) = V(X) + V(Y) - 2Cov(X,Y)$.

Exemple

Lors d'une enquête, on a demandé aux personnes interrogées dans une entreprise leurs nombre d'enfants à charge X et leur revenus mensuels Y qu'ils perçoivent en centaines d'euros. Le tableau reporte les résultats en fréquence arrondies au millième.

X \ Y	$[0;8[$	$[8;15[$	$[15;25[$	$[25;30[$	$[35;40[$	$[40;45[$	$[45;50[$	*Total*
1	0.048	0.079	0.085	0.091	0.012	0	0	0.315
2	0.036	0.091	0.060	0.054	0.018	0	0	0.259
3	0.024	0.030	0.060	0.048	0.006	0	0	0.168
4	0.012	0.018	0.054	0.018	0.018	0.018	0.006	0.144
5	0.012	0.006	0.030	0.018	0.018	0.012	0.006	0.102
Total	0.134	0.224	0.289	0.229	0.072	0.030	0.012	1

- La fréquence de trois enfants est 0.168 soit 16.8%.
- La fréquence de la classe de revenus mensuels $[15;25[$ est 0.289 soit 28.9%
- La fréquence du couple $(3;[15;25[)$ (personnes ayant 3 enfants et moins de 2500€ revenus mensuels) est environ 0.060 soit 60%.
- On a $0.168 \times 0.289 \approx 0.048 \neq 0.060$. Donc les variables X et Y sont dépendantes.

Remarque : il est inutile de faire le même calcul pour les autres couples (x_i, y_j), dans la mesure où l'égalité "$F_{ij} = f_i \times g_j$" n'est pas vérifiée.

Statistique ou test χ^2

- Le test χ^2 est un test d'indépendance entre deux variables statistiques X et Y portant sur une même population.
  - Il mesure l'indépendance entre deux caractères (variables quantitatives).
  - il compare les effectifs observées O (résultent des données statistiques) avec les effectifs théoriques T (résultent de l'indépendance des deux variables).
  - il s'appuie sur un mécanisme de test d'hypothèse (l'hypothèse nulle notée H_0 est l'indépendance des variables).
  - Il nécessite des effectifs théoriques supérieurs strictement à 5 et l'effectif total supérieur strictement à 50.

Examples

- Est-ce que l'appartenance politique dépend du sexe ?
- Est-ce que les jeunes filles anorexiques ont plus de chances d'être enfants uniques que les autres ?
- Est-ce que les immigrés nord-africains de seconde génération sont plus susceptibles de consommer régulièrement de l'alcool (oui/non) que les immigrés de première génération ?

Mise en œuvre du test d'indépendance χ^2

La mise en œuvre du test χ^2 s'appuie sur l'hypothèse nulle H_0 (indépendance des variables) et scindé en plusieurs étapes :

1. **Détermination des effectifs observés O :** donnés dans l'énonce ou calculé principalement à partir des fréquences ou toute autre donnée.
2. **Détermination des effectifs théoriques T :** on utilise les distributions marginales de X et de Y (données observées) et on obtient le tableau de contingence (données théoriques) via la formule

$$T_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

   où
   - T_{ij} est l'effectif théorique du couple (x_i, y_j) ;
   - $n_{i\cdot}$ est l'effectif marginal de x_i ;
   - $n_{\cdot j}$ est l'effectif marginal de y_j ;
   - n est l'effectif total ($n = \sum_{i=1}^{m} n_{i\cdot} = \sum_{j=1}^{k} n_{\cdot j} = \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}$).

   On vérifie que les conditions d'application du test sont vérifiées (tous les effectifs théoriques supérieurs à 5 et l'effectif total est supérieur à 50).
3. **Calcul de l'écart :** on calcule l'écart par la formule

$$\chi^2_{obs} = \sum \frac{(\text{effectif observés - effectif théorique})^2}{\text{effectif théorique}} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{T_{ij}}$$

Remarque : Il existe une autre formulation moins connue de ce test : $\chi^2_{obs} = \sum \frac{O_{ij}^2}{T_{ij}} - n.$

Mise en œuvre du test d'indépendance χ^2 (suite)

4. **Détermination de la valeur du χ^2 :**
   - on détermine d'abord le nombre de degrés de liberté (noté ddl) : si n est le nombre de lignes et m est le nombre de colonnes (hors ligne et colonne des distributions marginales) dans la table de contingence, alors $ddl = (n-1)(m-1)$.
   - On lit la valeur du $\chi^2_{1-\alpha}(ddl)$ dans la table en fonction du seuil de risque α (ou bien du niveau de confiance $1-\alpha$).

ddl \ α	0.90	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.016	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.211	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.584	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	1.064	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	1.610	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	2.204	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	2.833	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	3.490	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	4.168	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	4.865	9.342	11.781	13.442	15.987	18.397	21.161	23.209	29.588

5. **Conclusion sur l'hypothèse d'indépendance :**
   - $\chi^2_{obs} < \chi^2_{1-\alpha}(ddl)$, on accepte l'hypothèse d'indépendance des variables au seuil α.
   - $\chi^2_{obs} \geq \chi^2_{1-\alpha}(ddl)$, on rejette l'hypothèse au risque $1-\alpha$.

Remarque : Le seuil α correspond au risque de rejeter l'hypothèse d'indépendance alors que celle-ci est vraie. Par contre, on ne connait pas le risque d'accepter cette hypothèse alors qu'elle est fausse.

Application numérique

Une promotion de 150 étudiants en BUT-Techniques de Commercialisation est répartie en cinq groupes G_i, $1 \leq i \leq 5$. On souhaite savoir si le fait pour un étudiant de valider son année est indépendante du groupe d'appartenance avec un niveau de confiance 98% (ou seuil de risque 2%). Le tableau reporte les résultats où X est la variable correspondant aux groupes et Y est la variable correspondant à la validation de l'année.

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
V	22	20	24	20	25	111
$\bar{V}$	6	17	5	6	5	39
TOTAL	28	37	28	26	30	150

Réponse : on reprend les étapes de mise en œuvre du test χ^2.

1. **Détermination des effectifs observés O :** voir le tableau ci-dessus.
2. **Détermination des effectifs théoriques T :** on reprend le tableau uniquement avec les effectifs marginaux et on le complète sous l'hypothèse d'indépendance avec la relation $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$.

Y \ X	G_1	G_2	G_3	G_4	G_5	TOTAL
V	20.72	27.38	20.72	19.24	22.20	111
$\bar{V}$	7.28	9.62	7.28	6.76	7.8	39
TOTAL	28	37	28	26	30	150

Exemple : $T_{11} = \frac{n_{1.} \times n_{.1}}{n} = \frac{111 \times 28}{150} \approx 20.72$

D'après le tableau, les conditions d'application du test χ^2 sont vérifiées : tous les effectifs théoriques sont supérieurs à 5 et l'effectif total est supérieur à 50.

3. **Calcul de l'écart :** on a

$$\chi^2_{obs} = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \frac{(22-20.72)^2}{20.72} + \frac{(20-27.38)^2}{27.38} + \frac{(24-20.72)^2}{20.72} + \frac{(20-19.24)^2}{19.24} + \frac{(25-22.20)^2}{22.20}$$

$$+ \frac{(6-7.28)^2}{7.28} + \frac{(17-9.62)^2}{9.62} + \frac{(5-7.28)^2}{7.28} + \frac{(6-6.76)^2}{6.76} + \frac{(5-7.8)^2}{7.8} \approx 10.662$$

4. **Détermination de la valeur du χ^2 :**
   - Le tableau de contingence est composé de 2 lignes et 5 colonnes, donc $ddl = (2-1)(5-1) = 4$
   - Pour $\alpha = 98\%$ et $ddl = 4$, la valeur du $\chi^2_{1-\alpha}(ddl)$ dans la table est $\chi^2_{1-\alpha}(ddl) = \chi_{0.02}(4) = 11.668$.

5. **Conclusion sur l'hypothèse d'indépendance :** comme $\chi^2_{obs} < \chi^2_{1-\alpha}(ddl)$, alors on accepte l'hypothèse d'indépendance de X et de Y, et on en déduit que l'appartenance au groupe et la validation de l'année sont indépendantes au seuil de 2%.

Remarque : Pour un niveau de confiance $\alpha = 95\%$, on a $\chi^2_{1-\alpha}(ddl) = \chi_{0.05}(4) = 9.488$. Comme $\chi^2_{obs} > \chi^2_{1-\alpha}(ddl)$, alors on rejette l'hypothèse d'indépendance de X et de Y.

Chapitre 3 : Ajustement ou régression linéaire

1. Définition de nuages de points d'une série statistique
2. Ajustement par la méthode des moindres carrées
3. Coefficient de corrélation linéaire
4. Exercices d'application

$\leadsto$ **Mise en application en TD n° 2 et TP n° 2.**

Ajustement linéaire ou droite de régression

Définition

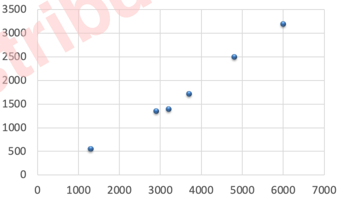
Soit (X, Y) une série statistique à deux variables quantitatives observées sur une même population de taille n dont les valeurs associées du i^{eme} individus sont (x_i, y_i), $1 \leq i \leq n$. On appelle **nuage de points** associé à la série (X, Y) l'ensemble des points du plan (muni d'un repère orthogonal $(O, \vec{i}, \vec{j})$) de coordonnées (x_i, y_i), $1 \leq i \leq n$.

Exemple 1

Données du constructeur de voitures

Régions	X	Y
1	1300	550
2	3200	1400
3	6000	3200
4	2900	1350
5	3700	1715
6	4800	2500

Nuage de points associé à la série (X, Y)



En reliant les différents points, on en déduit que la courbe d'ajustement du nuage de points semble être une droite linéaire.

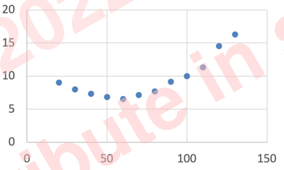
Exemple 2

On s'intéresse à la consommation du gasoil d'une voiture en fonction de sa vitesse moyenne par tranche de $20\,km/h$. On note X la vitesse moyenne de la voiture et Y la consommation du gasoil.

Données

X	Y
20	9
30	8
40	7.3
50	6.8
60	6.5
70	7.1
80	7.7
90	9.1
100	10
110	11.3
120	14.5
130	16.3

Nuage de points associé à la série (X, Y)



En reliant les différents points, on en déduit que la courbe d'ajustement du nuage de points semble être une parabole.

Dans la suite, on ne s'intéresse qu'aux ajustements dont la courbe est une droite (la fonction associée est une fonction affine ou linéaire) : on parle d'**ajustement linéaire**.

Ajustement par la méthode des moindres carrés (ou la méthode Mayer)

Principe de la méthode des moindres carrés :

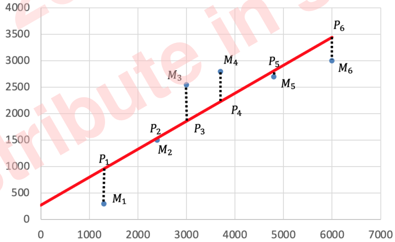
- Il consiste à déterminer la droite de la forme $D(Y/X)$: $y = ax + b$ telles que les distances prises entre chaque point du nuage et la droite soient les plus petites possibles.
- Pour mesurer l'écart entre la droite et les points du nuage, on utilise la somme des carrés des écarts (SCE) verticaux entre les points P_i et M_i, c'est-à-dire :

$$M_1P_1^2+M_2P_2^2+M_3P_3^2+M_4P_4^2+M_5P_5^2+M_6P_6^2$$

Autrement dit, on minimise la somme des carrés des écarts verticaux (SCE) :

$$SCE = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

où x_i et y_i sont les données, et a et b sont les inconnues



- Par un calcul simple, on en déduit qu'il existe une unique droite du plan, telle que $SCE = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$ soit minimale, d'équation $y = ax + b$ où

$$a = \frac{Cov(X,Y)}{V(X)} \ \text{(pente)}, \qquad b = \bar{y} - a\bar{x} \ \text{(ordonné)}. \qquad (2$$

Exemple

Reprenons l'exemple du constructeur de voitures.

- Déterminer la droite de régression $D(Y/X)$: $y = ax + b$. D'après le calcul effectué dans les sections précédentes, on a $Cov(X,Y) \approx 1257625$ et $V(X) = \frac{1300^2+3200^2+6000^2+2900^2+3700^2+4800^2}{6} - 3650^2 = 2189166.667$, si bien que $a = 0.5745$ et $b \approx -311.0062$.
  Finalement, la droite de régression est : $y = 0.5745x - 311.0062$.
- Estimer le nombre d'options vendues pour 13000 voitures :

$$\hat{y}(13000) = 0.5745 \times 13000 - 311.0062 \approx 6582,9938 \approx 6583 \text{ options vendues}$$

Propriétés

- La droite de régression de Y en X passe par le point moyen $(\bar{x}, \bar{y})$ ou le **centre de gravité** $G(\bar{x}, \bar{y})$ du nuage des observations.
- La droite de régression de Y en X passe par le point $(0, b)$.

Remarque : Pour obtenir la droite de régression de X en Y, il suffit simplement de permuter les rôles de X et de Y. On a ainsi : $D' = D(X/Y)$: $x = a'y + b'$ où

$$a' = \frac{Cov(X,Y)}{V(Y)} \quad \text{et} \quad b' = \bar{X} - a'\bar{Y}.$$

Corrélation et coefficient de corrélation linéaire

- **Corrélation :** deux variables X et Y sont **corrélées** si l'on observe une dépendance, une relation entre les deux.
- **Corrélation ou causalité ?** une erreur courante permet de dire "X et Y sont corrélées", alors X cause Y. On confond ainsi la corrélation et la causalité parce qu'en réalité, il se pourrait aussi que
  - Y cause X.
  - X et Y ont une cause commune ou cause cachée C.
  - X et Y soient accidentellement liés mais n'ai aucun lien de causalité.

Exemples

- Nombre de cheveux d'un homme a tendance à diminuer avec l'âge.
- Poids et taille d'un nouveau née.
- Consommation de gasoil et vitesse d'une voiture.
- Chiffre d'affaires et les charges de publicités d'une entreprise.

Définition

Le **coefficient de corrélation linéaire** $r(X,Y)$ entre les variables X et Y est une mesure de la qualité d'ajustement ou le degré d'indépendance entre les deux variables (si la variance de X et celle de Y sont non nuls) :

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(x)V(Y)}} \quad \text{où } V(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2, \quad V(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2.$$

Exemple : reprenons l'exemple 1. du constructeur de voitures

Avec les notations ci-dessus, on a : $Cov(X,Y) = 1257625$ et

$$V(X) = \frac{1300^2 + 3200^2 + 6000^2 + 2900^2 + 3700^2 + 4800^2}{6} - 3650^2 = 2189166.667$$

$$V(Y) = \frac{550^2 + 1400^2 + 3200^2 + 1350^2 + 1715^2 + 2500^2}{6} - 1777.5^2 \approx 730170.1389.$$

D'où $r(X,Y) = \dfrac{1257625}{\sqrt{2189166.67 \times 730170.1389}} \approx 0.9947$

$\rightsquigarrow$ le nombre de voitures vendues X et le nombre d'options vendues Y sont fortement corrélées (Autrement dit ; X explique Y à 99.47%).

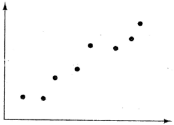
Remarques :

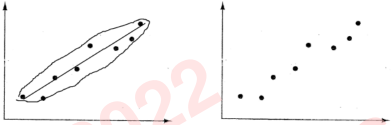
- $r(X,Y)$ est un coefficient sans unité (dimension) vérifiant : $-1 \leq r(X,Y) \leq 1$; $r(X,Y) = r(Y,X)$ et $r(X,X) = 1$.
- $r(X,Y) \approx 1$: une corrélation positive et présence d'une relation linaire entre X et Y (même sens de variation).
- $r(X,Y) \approx -1$: une corrélation négative et présence d'une relation linaire entre X et Y (différent sens de variation).
- $r(X,Y) \approx 0$: absence de relation linéaire entre X et Y mais peut-être une autre forme de liaison non-linéaire.
- La corrélation est dite forte si les points du nuage sont très proches de la droite d'ajustement (boîte très étroite).
- La corrélation est dite faible si les points du nuage sont assez proches de la droite d'ajustement (boîte assez étroite).

Évaluation graphique du coefficient de corrélation $r(X, Y)$

- corrélation positive forte

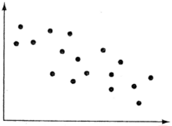
nuage ou boîte proche de la droite de régression

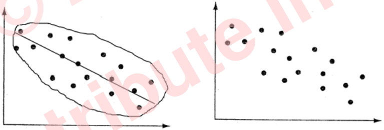




- corrélation négative faible

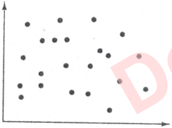
nuage ou boîte éloigné de la droite de régression

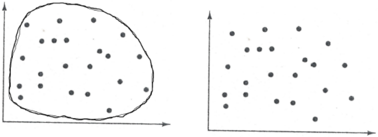




- pas de corrélation
  pas de droite de régression

nuage sans 'direction' ou boîte 'carrée' ou horizontale
pas de droite de régression





Exemples : voir TD3

Exercice 1

Un chercheur universitaire souhaite mesurer l'influence des dépenses de publicité sur la vente des boîtes de conserve. Les informations sont récoltées dans le tableau ci-dessus :

Ventes (10^3 boîtes)	10	25	30	35	45	55	70	75
Dépenses de publicité (10^3€)	2	6	8	10	12	17	22	24

1. Réaliser le nuage de points en plaçant le point moyen.
2. Calculer le coefficient de corrélation linéaire. En déduire que l'ajustement linéaire se justifie dans cette situation.
3. Déterminer l'équation de la droite des moindres carrées.
4. Grâce à cette équation, déterminer une estimation de ventes prévisible pour des dépenses de publicité de 20000€.

Chapitre 4 : séries chronologiques

1. Introduction et objectifs de l'étude des chronologiques
2. Modèle additif
3. Modèle multiplicatif

$\rightsquigarrow$ **Mise en application en TD n° 3 et TP n° 3.**

Introduction et objectif de l'étude des séries chronologiques

Définition

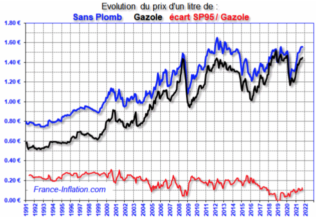
- Une **série chronologique (ou chronique ou temporelle)** est une série statistique de variables quantitatives ayant la particularité de décrire un phénomène dans le temps t qui peut être selon les cas, la seconde, la minute, l'heure, le jour, le mois, l'année, ...
- S'il y a n observations, celles-ci sont numérotées par un indice t et la série se note $(y_t)_{t\in\{1,\dots,n\}}$.
- **Applications :** économie, finance, démographie, médecine, énergie, ...

Objectif de l'étude des séries chronologiques

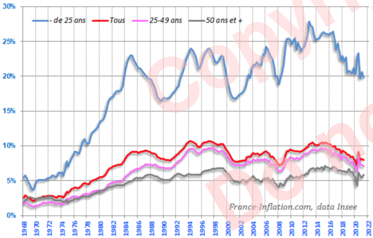
- Description du passé : analyse d'un phénomène temporel en mettant en évidence essentiellement la tendance générale et les fluctuations saisonnières ;
  - Quelle a été l'évolution générale ?
  - Y a-t-il eu des phénomènes saisonniers ? Peut-on les mesurer ?
  - Y a-t-il des ruptures (temporaires ou permanences) dans cette évolution ?
  - Dans quelle mesure une observation à un instant i est-elle liée aux observations passées ?
- Prévisions du futur : élaboration d'un modèle de prévision à court terme des valeurs non observées ;
  - Quelles valeurs peut-on prévoir pour la variable sur la période à venir ?
  - Peut-on élaborer un modèle permettant de faire de la prévision à court terme ?

Exemples tirés de "France-Inflation & CGEDD"

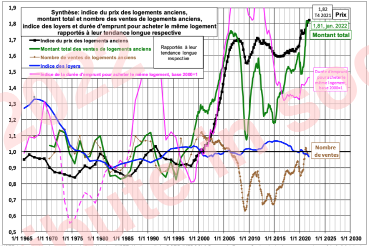
- Évolution du prix d'un litre du pétrole



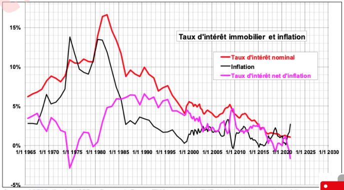
- Évolution du taux de chômage en France



- Évolution du marché immobilier : actualisations à 2022



- Évolution du taux d'intérêt immobilier



Exemples

Constructeur français de voitures

On s'intéresse au nombre de voitures vendues par le constructeur français, au cours de chacun 16 dernières trimestres depuis Covid-19.

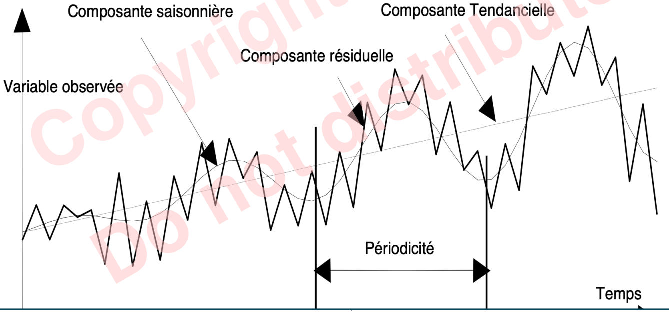
A \ T	T_1	T_3	T_3	T_4
2018	75	152	43	29
2019	87	158	49	33
2020	96	164	54	37
2021	105	170	58	41
2022	$\cdots$	$\cdots$	$\cdots$	$\cdots$

- Comment peut-on utiliser ces données pour prédire convenablement combien de voitures seront vendues durant les quelques prochains trimestres ?
- Le volume général des ventes est-il en progression ?
- Y a-t-il un phénomène saisonnier suffisamment important pour qu'on doive en tenir compte.

Propriétés

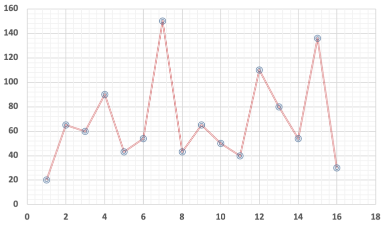
Une série chronologique Y_t peut se décomposer en plusieurs composantes :

- une **composante tendancielle (ou trend)** T_t, qui décrit le mouvement à long terme de la grandeur et traduit le comportement général de la série(croissance, décroissance,. ..).
- une **composante saisonnière** S_t, qui donne la tendance périodique du caractère étudié dépendant des saisons (production agricole, ...)
- une **composante résiduelle (ou aléatoire ou accidentelle)** ε_t qui prend en compte les aléas (Covid-19, grèves,catastrophe naturelle, crash financier, Invasion de l'Ukraine, ...)

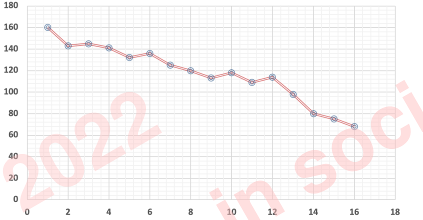


Exemples : on représente quatre séries chronologiques résultant de l'étude du chiffre d'affaires d'une entreprise sur les quatre trimestres de quatre année.

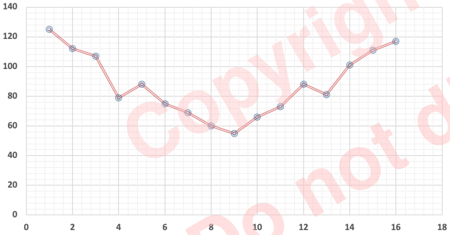
- Sans tendance, ni saisonnalité



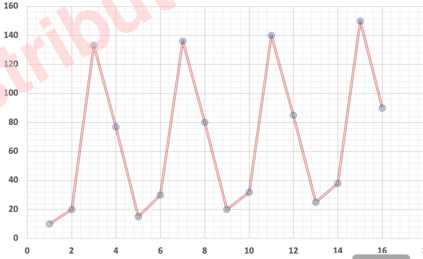
- Tendance linéaire, sans saisonnalité



- Tendance non linéaire, sans saisonnalité



- Tendance non linéaire, saisonnalité



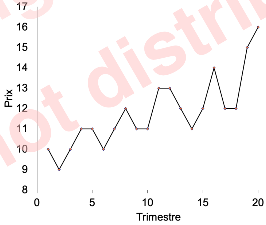
Exemple : détermination des composantes de la série chronologique

On s'intéresse au prix P_t d'un article en fonction du temps t. Le tableau reporte les prix par trimestre entre 2018 et 2022.

Date	Prix €
T1 2018	10
T2 2018	9
T3 2018	10
T4 2018	11
T1 2019	11
T2 2019	10
T3 2019	11
T4 2019	12
T1 2020	11
T2 2020	11
T3 2020	13
T4 2020	13
T1 2021	12
T2 2021	11
T3 2021	12
T4 2021	14
T1 2022	12
T2 2022	12
T3 2022	15
T4 2022	16

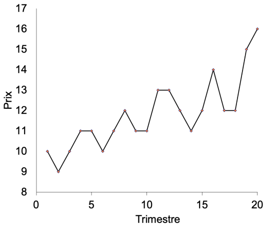
Trimestre \ Année	2018	2019	2020	2021	2022
1^{er} trimestre	10	11	11	12	12
$2^{ème}$ trimestre	9	10	11	11	12
$3^{ème}$ trimestre	10	11	13	12	15
$4^{ème}$ trimestre	11	12	13	14	16

- Série chronologique des prix P_t en fonction du temps t (trimestre)

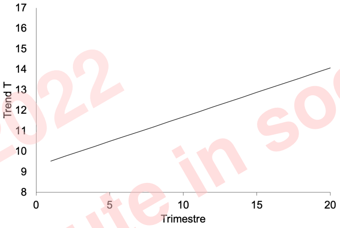


Les composantes de la série P_t

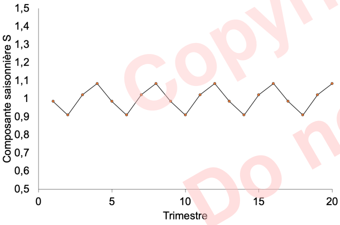
Série initial P_t



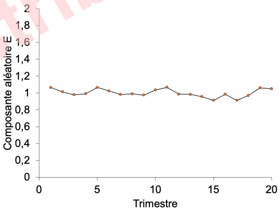
Tendance ou trend T_t



Composante saisonnière S_t



Composante aléatoire ε_t



Prendre en main les données chronologiques

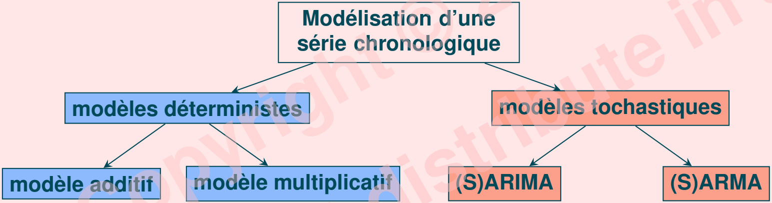
Les principales étapes simplifiées sont les suivantes :

1. Décrire les données chronologiques ou temporelles :
   - Faire un constat sur les moyennes, les dispersions, les relations entre les moyennes des variables (ratios, ...).
   - Étudier la dynamique : évolution des valeurs observées et taux de croissance.
   - Traitement des données : tableaux de synthèse et graphiques.
2. Choisir la périodicité à étudier :
   - Trop longue : les facteurs extrêmes modifient les effets des variables.
   - Trop courte : les coefficients sont mal estimés, instables, ...
   - Intégration de la saisonnalité : mensualité, hebdomadaire, dé-saisonnalité ou intégrer une variable d'activité ou auxiliaire.
3. Choisir le niveau d'agrégation des données ou modèle de composition : additif ou multiplicatif :
   - Problématique de l'hétérogénéité.
   - Comportements en super/hypermarchés sont-ils identiques ?
   - Identification de la tendance et calcul des coefficients de variations saisonnières.
4. Prévision à court terme et évaluation de la tendance à un moment futur :
   - La reconstitution de la série avec ses trois composantes modélisées.
   - une méthode de prédiction par lissages exponentiels.

Modélisation d'une série chronologique

Définition

- Un **modèle** est une présentation simplifiée de la réalité qui vise à traduire les mécanismes de fonctionnement d'un phénomène et permet de mieux les comprendre.
- Un **modèle de série chronologique** est une équation précisant la façon dont les composantes s'articulent les unes par rapport aux autres pour constituer la série chronologique.

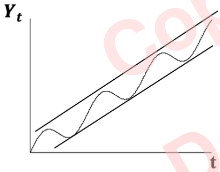


Dans ce cours, nous n'étudierons que les modèles déterministes.

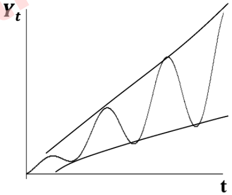
Remarque : Un modèle peut être meilleur qu'un autre pour décrire la réalité et bien sûr, plusieurs questions se posent alors : comment mesurer cette qualité et diagnostiquer un tel modèle

Présentation des modèles déterministes

- **Modèle additif :** la saisonnalité s'additionne à la tendance ; $Y_t = T_t + S_t + \varepsilon_t$ pour tout $t = 1, \cdots, n$.
  $\rightsquigarrow$ l'amplitude des variations est constante autour de la tendance.
  $\rightsquigarrow$ les nuages des points par période sont sensiblement les mêmes (superposition).
  $\rightsquigarrow$ les 2 droites tracées sont à peu près parallèles.
  $\rightsquigarrow$ l'étendue sur chaque période est sensiblement constante.
- Représentation graphique



- **Modèle multiplicatif :** la saisonnalité est proportionnelle à la tendance ;
  $Y_t = T_t \times S_t + \varepsilon_t$ pour tout $t = 1, \cdots, n$.
  $\rightsquigarrow$ l'amplitude des variations est variable autour de la tendance.
  $\rightsquigarrow$ les nuages des points par période sont sensiblement de plus en plus grands/petits (superposition).
  $\rightsquigarrow$ les 2 droites tracées ne sont pas parallèle
  $\rightsquigarrow$ l'étendue sur chaque période est de plus grande/petite.
- Représentation graphique



Exemple 1 : observation traitée selon un modèle additif

On s'intéresse au nombre de voitures vendues par le constructeur français sur chacun des quatre trimestres (T) de l'année depuis 2018 jusqu'à 2021.

- Le tableau suivant reporte les valeurs (en centaines de voitures vendues).

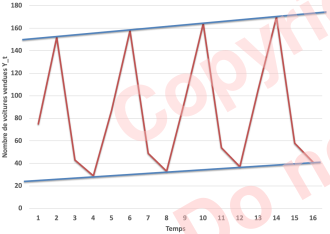
A \ T	T_1	T_3	T_3	T_4
2018	75	152	43	29
2019	87	158	49	33
2020	96	164	54	37
2021	105	170	58	41

- En calculant l'étendue sur chaque année :

$$E(A_1) = 152 - 29 = 123;\ E(A_2) = 125$$
$$E(A_3) = 164 - 37 = 127;\ E(A_4) = 129$$

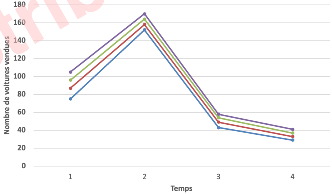
$\rightsquigarrow$ l'étendue sur chaque année semble donc être sensiblement la même.

- En représentant les nuages de points :



$\rightsquigarrow$ les droites des extrêmes semblent parallèles.

- En superposant les nuages de points par année :



$\rightsquigarrow$ les nuages de points semblent se superposer.

Modèle additif : détermination de la tendance par la méthode des moindres carrés

- Méthode des moindres carrés ou ajustement linéaire, (voir le chapitre 3) : $y = ax + b$ où $a = \frac{Cov(X,Y)}{V(X)}$ et $b = \bar{Y} - a\bar{X}$.
- Exemple : on reprend les données du constructeur français de voitures. On notera par X le rang du trimestre et Y le nombre de centaines de voitures vendues.

$X = t$	Y
1	75
2	152
3	43
4	29
5	87
6	158
7	49
8	33
9	96
10	164
11	54
11	37
13	105
14	170
15	58
16	41

Pour déterminer la tendance, on suit les étapes ci-dessous :

- Calcul de la moyenne de X et celle de Y : $\bar{X} = 8.5$ et $\bar{Y} = 84.4375$.
- Calcul de la variance de X : $V(X) = 21.25$.
- Calcul de la covariance $Cov(X,Y) = \overline{XY} - \bar{X} \times \bar{Y} = -3.53125$.
- Calcul de la pente $a = \frac{Cov(X,Y)}{V(X)} = \frac{-3.53125}{21.5} \approx -0.17$.
- Calcul du point d'origine $b = \bar{Y} - a\bar{X} = 85.9$.

$\rightsquigarrow$ la droite d'ajustement linéaire de la tendance est $y = -0.17x + 85.9$.

Modèle additif : détermination des coefficients saisonniers

- Les écarts saisonniers sont les différences entre les valeurs du caractère étudié Y et les valeurs de la tendance T.
- Le coefficient saisonnier est la moyenne des écarts saisonniers d'une même saison sur l'ensemble des périodes.
- Exemple : on reprend les données du constructeur français de voitures pour le modèle additif.

$X = t$	Y	$T = -0.17X + 85.9$	$S = Y - T$
1	75	85.7	-10.7
2	152	85.6	66.4
3	43	85.4	-42.4
4	29	85.2	-56.2
5	87	85.1	1.9
6	158	84.9	73.1
7	49	84.7	-35.7
8	33	84.5	-51.5
9	96	84.4	11.6
10	164	84.2	79.8
11	54	84.0	-30.0
11	37	83.9	-46.9
13	105	83.7	21.3
14	170	83.5	86.5
15	58	83.4	-25.4
16	41	83.2	-42.2

On en déduit les coefficients saisonniers de chaque trimestre :

- pour le coefficient saisonnier du 1^{er} trimestre, on calcule la moyenne des écarts saisonniers des trimestres 1, 5, 9 et 13 : $S_1 = \frac{-10.7+1.9+11.6+21.3}{4} = 6.025.$
- pour le coefficient saisonnier du $2^{ème}$ trimestre : $S_2 = \frac{66.4+73.1+79.8+86.5}{4} = 76.45.$
- pour le coefficient saisonnier du $3^{ème}$ trimestre : $S_3 = \frac{-42.4-35.7-30-25.4}{4} = -33.375.$
- pour le coefficient saisonnier du $4^{ème}$ trimestre : $S_4 = \frac{-56.2-51.5-46.9-42.2}{4} = -49.20$

$\rightsquigarrow \frac{S_1+S_2+S_3+S_4}{4} \approx 0$ (Toujours vraie).

Exemple 1 : observation traitée selon un modèle multiplicatif

On s'intéresse au nombre de voitures vendues par le constructeur français sur chacun des quatre trimestres (T) de l'année depuis 2018 jusqu'à 2021.

- Le tableau suivant reporte les valeurs (en centaines de voitures vendues).

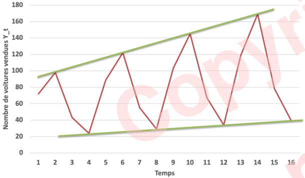
A \ T	T_1	T_3	T_3	T_4
2018	72	98	43	24
2019	89	122	55	29
2020	103	145	67	34
2021	119	169	78	39

- En calculant l'étendue sur chaque année :

$$E(A_1) = 98 - 24 = 74;\ E(A_2) = 39$$
$$E(A_3) = 154 - 34 = 111;\ E(A_4) = 130$$

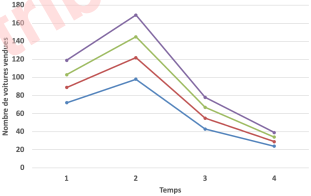
⇝ l'étendue sur chaque année semble donc être sensiblement la même.

- En représentant les nuages de points :



⇝ les droites des extrêmes ne sont pas parallèles.

- En superposant les nuages de points par année :



⇝ les nuages de points ne superposent (plus en plus grands)

Modèle multiplicatif : détermination de la tendance par la méthode des moindres carrés

- Méthode des moindres carrés ou ajustement linéaire, (voir le chapitre 3) : $y = ax + b$ où $a = \frac{Cov(X,Y)}{V(X)}$ et $b = \bar{Y} - a\bar{X}$.
- Exemple : on reprend les données du constructeur français de voitures. On notera par X le rang du trimestre et Y le nombre de centaines de voitures vendues.

$X = t$	Y
1	72
2	98
3	43
4	24
5	89
6	122
7	55
8	29
9	103
10	145
11	67
11	34
13	119
14	169
15	78
16	39

Pour déterminer la tendance, on suit les étapes ci-dessous :

- Calcul de la moyenne de X et celle de Y : $\bar{X} = 8.5$ et $\bar{Y} = 80.375$.
- Calcul de la variance de X : $V(X) = 21.25$.
- Calcul de la covariance $Cov(X,Y) = \overline{XY} - \bar{X} \times \bar{Y} = 36.5625$.
- Calcul de la pente $a = \frac{Cov(X,Y)}{V(X)} = \frac{36.5625}{21.5} \approx 1.72$.
- Calcul du point d'origine $b = \bar{Y} - a\bar{X} = 65.8$.

$\rightsquigarrow$ la droite d'ajustement linéaire de la tendance est $y = 1.72x + 65.8$.

Modèle multiplicatif : détermination des coefficients saisonniers

- Les rapports saisonniers sont les quotients entre les valeurs du caractère étudié Y et les valeurs de la tendance T.
- Le coefficient saisonnier est la moyenne des rapports saisonniers d'une même saison sur l'ensemble des périodes.
- Exemple : on reprend les données du constructeur français de voitures pour le modèle multiplicatif

$X = t$	Y	$T = 1.72X + 65.8$	$S = \frac{Y}{T}$
1	72	67.5	1.1
2	98	69.2	1.4
3	43	71.0	0.6
4	24	72.2	0.3
5	89	74.4	1.2
6	122	76.1	1.6
7	55	77.8	0.7
8	29	79.6	0.4
9	103	81.3	1.3
10	145	83.0	1.7
11	67	84.7	0.8
11	34	86.4	0.4
13	119	88.2	1.3
14	169	89.9	1.9
15	78	91.6	0.9
16	39	93.3	0.4

On en déduit les coefficients saisonniers de chaque trimestre :

- pour le coefficient saisonnier du 1^{er} trimestre, on calcule la moyenne des rapports saisonniers des trimestres $1, 5, 9$ et 13 : $S_1 = \frac{1.1+1.2+1.3+1.3}{4} = 1.225$.
- pour le coefficient saisonnier du $2^{\text{ème}}$ trimestre : $S_2 = \frac{1.4+1.6+1.7+1.9}{4} = 1.65$.
- pour le coefficient saisonnier du $3^{\text{ème}}$ trimestre : $S_3 = \frac{0.6+0.7+0.8+0.9}{4} = 0.75$.
- pour le coefficient saisonnier du $4^{\text{ème}}$ trimestre : $S_4 = \frac{0.3+0.4+0.4+0.4}{4} = 0.375$

$\rightsquigarrow \frac{S_1+S_2+S_3+S_4}{4} \approx 1$ (Toujours vraie).

Prévision et évaluation de la tendance à un moment futur

On souhaite prévoir le nombre de voitures que le constructeur pourra vendre au $3^{\text{ème}}$ trimestre de l'année 2022 en utilisant les différents modèles. Ce semestre est le trimestre de rang 19.

Modèle additif

- La tendance pour le trimestre 19 est $T_{19} = -0.17 \times 19 + 85.9 \approx 82.7$.
- La prévision à la date future $t = 19$, correspondant au 19 trimestre :
  $\hat{y}_{19} = T_{19} + S_3$.
  ↫ Ici, on utilise le coefficient saisonnier S_3 car on cherche à prévoir le nombre de voitures à vendre d'un $3^{\text{ème}}$ trimestre.
- $\hat{y}_{19} = 82.7 + (33.375) = 49.325 \approx 93.30$ (exprimé en centaines $\times 100$).

⇝ On peut ainsi prévoir que le constructeur vendra 4933 voitures au $3^{\text{ème}}$ trimestre de l'année 5

Modèle multiplicatif

- La tendance pour le trimestre 19 est $T_{19} = 1.72 \times 19 + 65.8 \approx 98.5$.
- La prévision à la date future $t = 19$, correspondant au 19 trimestre :
  $\hat{y}_{19} = T_{19} \times S_3$.
  ↫ Ici, on utilise le coefficient saisonnier S_3 car on cherche à prévoir le nombre de voitures à vendre d'un $3^{\text{ème}}$ trimestre.
- $\hat{y}_{19} = 98.5 \times (0.75) = 73.875 \approx 73.88$ (exprimé en centaines $\times 100$).

⇝ On peut ainsi prévoir que le constructeur vendra 7388 voitures au $3^{\text{ème}}$ trimestre de l'année 5.

Prendre en main les données chronologiques

Les principales étapes simplifiées sont les suivantes :

1. Décrire les données chronologiques ou temporelles :
   - Faire un constat sur les moyennes, les dispersions, les relations entre les moyennes des variables (ratios, ...).
   - Étudier la dynamique : évolution des valeurs observées et taux de croissance.
   - Traitement des données : tableaux de synthèse et graphiques.
2. Choisir la périodicité à étudier :
   - Trop longue : les facteurs extrêmes modifient les effets des variables.
   - Trop courte : les coefficients sont mal estimés, instables, ...
   - Intégration de la saisonnalité : mensualité, hebdomadaire, dé-saisonnalité ou intégrer une variable d'activité ou auxiliaire.
3. Choisir le niveau d'agrégation des données ou modèle de composition : additif ou multiplicatif :
   - Choix du modèle adapté.
   - Comportements en super/hypermarchés sont-ils identiques ?
   - Identification de la tendance et calcul des coefficients de variations saisonnières.
4. Prévision à court terme et évaluation de la tendance à un moment futur :
   - La reconstitution de la série avec ses trois composantes modélisées.
   - une méthode de prédiction par lissages exponentiels.

Fin du cours ! ! !

Source de l'image (https://fr.123rf.com)

Ⓢ Contrôle continu, le 08 avril 2022.