Échantillonnage et Estimation

Mohamed El Omari
Enseignant Chercheur,
Spécialité: Statistique et Probabilités

Ancien Inspecteur Pédagogique

Faculté Polydisciplinaire de Sidi Bennour.

December 24, 2021

Outline

1. Rappel: Statistique descriptive

1.1. Vocabulaire

- Population, échantillon et variable statistique.
- Types de variable statistique.
- Effectif, fréquence et pourcentage.
- Représentation graphique des données statistiques.

Exercice 1: Considérons les données graphiques suivantes :

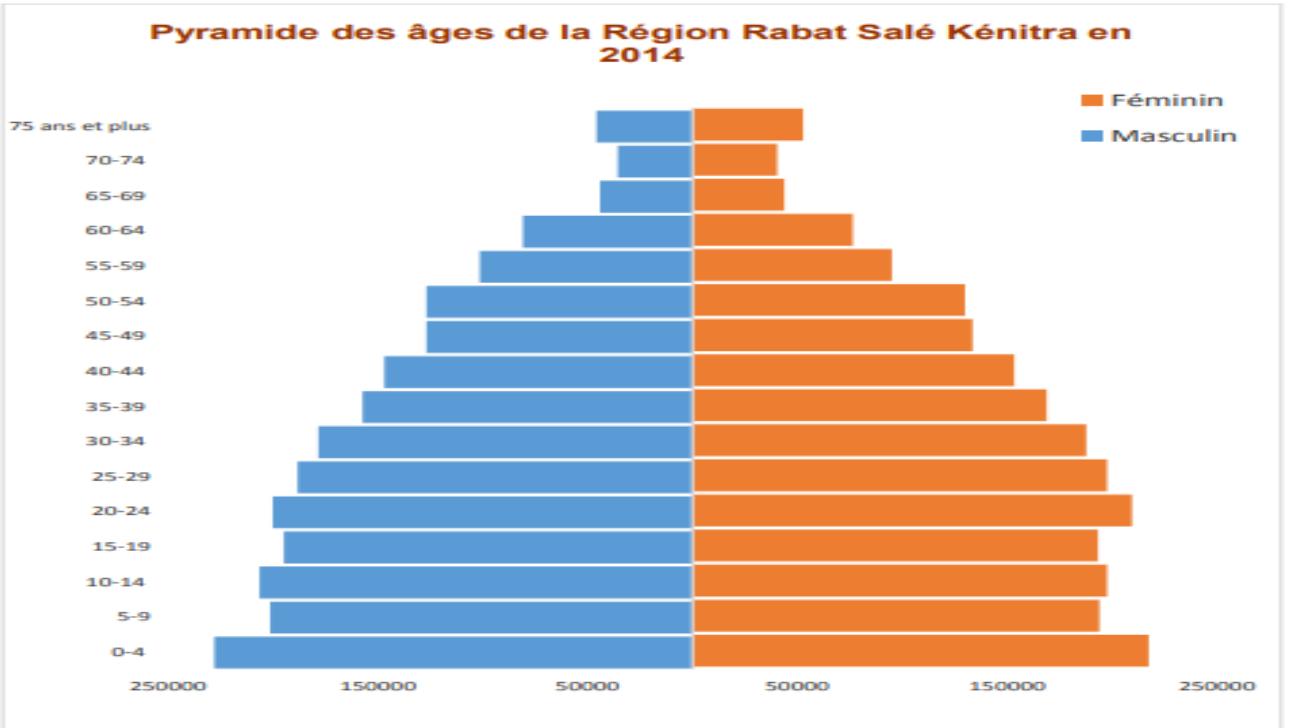


Figure 1: Source: www.hcp.ma/region-rabat/attachment/775960/

1. Déterminer la population étudiée.
2. Identifier la nature des caractères étudiés.

Exercice 2: Pour évaluer les performances des élèves d'une classe collégiale en mathématiques, une épreuve a été faite. On a obtenu la série suivante:
8 - 11 - 13 - 5 - 8 - 14 - 6 - 12 - 5 - 10
16 - 7 - 12 - 13 - 8 - 13 - 8 - 7 - 13 - 13
9 - 17 - 10 - 13 - 6 - 13 - 7 - 14

1. Déterminer la population étudiée.
2. Quelle est la variable statistique? De quel type est-elle? Comment peut-on organiser les données?

Exercice 3: On interroge 50 personnes sur leur dernier diplôme obtenu. La codification a été faite selon le tableau suivant:

Dernier diplôme obtenu	**Modalité x_j**
Sans diplôme	Sd
Primaire	P
Secondaire	Se
Supérieur non-universitaire	Su
Universitaire	U

1. Déterminer la population, le caractère étudié ainsi que la nature de ce caractère.

2 compléter le tableau suivant.

Modalité x_j	**Effectif**	**Fréquence**	**Pourcentage**
Sd	·	0,08	·
P	11	·	·
Se	·	·	28
Su	9	·	·
U	·	0,24	·
Total	50	·	·

1.2. Paramètres de position

- Le mode
- La moyenne
- La médiane
- Le quantile

1.3. Paramètres de dispersion

- L'étendue
- L'intervalle interquartile
- La variance
- L'écart-type

Exercice 4: Une population de ménages a été répartie en fonction du nombre de parts familiales permettant le calcul de l'impôt sur le revenu.

Nombre de Parts	1	1,5	2	2,5	3	3,5	4	4,5
Nombre de Ménages	48	58	136	184	210	122	62	12

1. Quel est le caractère étudié et quelles sont les modalités?
2. Déterminer l'étendue et le mode de cette série statistique.
3. Donner une représentation graphique de cette population.
4. Calculer la médiane, la moyenne et l'écart-type de cette variable.

Pourquoi faire l'échantillonnage?

Pour des considérations pratiques, la variable d'intérêt n'est pas observée sur l'ensemble de la population :

- Effectuer un recensement coûte cher, et suppose de disposer d'une base de sondage donnant la liste de l'ensemble des individus de la population.
- Dans certaines situations il est impossible de faire un recensement.

- Même dans le cas d'un recensement traditionnel, l'ensemble des données recueillies est rarement exploité.
- Augmenter la taille d'un questionnaire augmente le fardeau de réponse, et diminue les taux de réponse.
- De façon générale, la non-réponse diminue la taille de l'échantillon effectivement observé.

2. Introduction à la théorie de l'échantillonnage

2.1. Vocabulaire

2.2. Méthodes d'échantillonnage

2.1. Vocabulaire

- **Enquête:** Une enquête peut être n'importe quelle activité de collecte d'information organisée et méthodique à propos des caractéristiques des unités d'une population.
- **Recensement:** Enquête complète ou enquête exhaustive, c'est une enquête au cours de laquelle toutes les unités de base de la population sont observées.

- **Sondage:** Enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête au cours de laquelle seulement une partie des unités de base de la population sont observée.
- **Échantillon:** Ensemble des unités de base sélectionnées et réellement observées au cours d'un sondage.
- **Échantillonnage:** Ensemble des opérations qui permettent de sélectionner de façon organisée les éléments de l'échantillon.

- **Unité de base:** Unité d'échantillonnage ou unité de sondage, c'est l'élément pris en considération dans l'enquête.
- **Base de sondage:** Énumération ou présentation ordonnée de toutes les unités de base constituant la population.

- **Erreur d'échantillonnage:** Écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement comparable de la population.
- **Fraction ou taux de sondage:** Proportion des unités de la population qui font partie de l'échantillon. C'est le rapport entre la taille de l'échantillon n et la taille de la population N.

$$t = \frac{n}{N}$$

2.2. Méthodes d'échantillonnage

2.2.1. Méthodes d'échantillonnage probabilistes

2.2.2. Méthodes d'échantillonnage empiriques

2.2.1. Méthodes d'échantillonnage probabilistes

- Échantillonnage aléatoire et simple
- Échantillonnage systématique
- Échantillonnage aléatoire en grappes
- Échantillonnage aléatoire stratifié
- Échantillonnage à plusieurs degrés

Échantillonnage aléatoire et simple

- Les éléments de l'échantillon sont choisis aléatoirement (en utilisant par exemple une table de nombres aléatoires, un logiciel statistique, fonction **ALEA** Excel) à partir d'une liste énumérative de tous les éléments.
- Les individus de la population ont la même probabilité d'être retenu lors du tirage au sort.
- Favorise la représentativité (mais ne la garantit pas!).
- Simple mais peut être difficile d'utilisation et onéreux lorsqu'il n'existe pas de liste et qu'il faut la construire.

Échantillonnage systématique

- Les éléments de l'échantillon sont sélectionnés à intervalles réguliers.
- Le premier individu est choisi aléatoirement.
- On ne connaît pas la probabilité d'inclusion des individus de la population.
- La représentativité de l'échantillon n'est nullement garantie.

Échantillonnage aléatoire en grappes

- Choix aléatoire de grappes (sous-groupes de la population) au lieu d'unités.
- Utile lorsque les éléments sont naturellement groupés ou quand il n'est pas possible d'obtenir la liste de tous les éléments de la population cible.
- Économique en temps et en argent; moins exact cependant que l'aléatoire simple.

Échantillonnage aléatoire stratifié

- Population découpée en strates représentant certaines de ses caractéristiques.
- Les éléments sont choisis dans les strates à l'aide d'une technique d'échantillonnage probabiliste; le nombre d'éléments choisis dans les strates peut ou non représenter les proportions de la population.
- Méthode la plus raffinée; permet d'assurer une meilleure représentativité et de comparer les sous-groupes.

Échantillonnage aléatoire à plusieurs degrés

L'échantillonnage aléatoire à plusieurs degrés regroupe toute une série de plans d'échantillonnage caractérisés par un système ramifié et hiérarchisé d'unités.

2.2.2. Méthodes d'échantillonnage empiriques (non probabilistes)

- Échantillonnage accidentel (de convenance)
- Échantillonnage par choix raisonné
- Échantillonnage volontaire
- Échantillonnage par réseau (boule de neige)
- Échantillonnage par quotas

Échantillonnage de convenance

- Les éléments de l'échantillon sont choisis au fur et à mesure qu'ils se présentent, sans tri.
- Simple, rapide, peu coûteux mais offre le moins de garantie.

Échantillonnage par choix raisonné

- Choix des éléments basé sur le **jugement du chercheur** par rapport à leur caractère typique ou atypique.
- Permet d'étudier des phénomènes rares ou inusités; peu de représentativité de l'ensemble de la population.

Échantillonnage volontaire

- Les éléments de l'échantillon sont choisis sur une base volontaire.
- Peut offrir une meilleure représentativité si on sélectionne parmi les volontaires; certain biais du fait que les volontaires ont certains traits de caractère particuliers (par exemple, les timides sont moins portés à participer).

Échantillonnage par réseau

- Les éléments de l'échantillon sont choisis à travers des réseaux sociaux, d'amitiés.
- Exemple: Choisir quelques personnes correspondant au profil recherché et leur demander de nous donner des noms de personnes "similaires".

Échantillonnage par quotas

- Population découpée en strates représentant certaines de ses caractéristiques.
- Les éléments sont choisis dans les strates à l'aide d'une **technique d'échantillonnage non probabiliste**.
- Le nombre d'éléments choisis dans les strates représente les proportions de la population.

3. Les modèles de régression linéaire (simples et multiples)

3.1. Rappel: Représentation graphique de deux séries statistiques

3.2. Rappel: Comparaison de deux séries statistiques

3.3. La droite de régression linéaire

3.1. Représentation graphique de deux séries statistiques

Les graphiques:

- se lisent et sont compris rapidement,
- **montrent les faits les plus importants,**
- facilitent la compréhension des données,
- peuvent convaincre le lecteur,
- **aident le statisticien à comparer les séries statistiques,**
- aident le lecteur à se souvenir des données.

Les graphiques à discuter:

- Diagramme chronologique (ou linéaire par morceaux); Diagramme en bâtons.
- Histogramme.
- Diagramme à barres; Diagramme à barres empilées.
- Diagramme circulaire (ou diagramme en secteurs).
- Boîte moustaches (ou diagramme en boîte).
- Nuage de points.

3.2. Rappel: Comparaison de deux séries statistiques

- Pour comparer deux séries statistiques on compare leur paramètres de position et de dispersion.
- L'écart-type, malgré sa pertinence dans la mesure de la dispersion d'une série statistique, possède un inconvénient majeur:
  - **(a)** Il est exprimé dans l'unité de la variable à laquelle il se rapporte.
  - **(b)** Il est alors impossible de comparer les dispersions de deux séries statistiques ayant un lien entre elles et dont les valeurs s'expriment dans des unités différentes.

- Pour comparer la dispersion de deux séries qui ne sont pas exprimées dans les mêmes unités, on utilise le coefficient de variation noté **CV**.
- Le **CV** d'une série statistique $\{x_1, \cdots, x_n\}$ de moyenne $\overline{x}$ et d'écart type s_x est défini par

$$\mathbf{CV} = \frac{s_x}{\overline{x}}.$$

- Une pratique empirique courante est de considérer que la série possède une variabilité significative si $\mathbf{CV} > 0.15$.
- Si $\mathbf{CV} \leq 0.15$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de toute la série.

Le coefficient de corrélation de deux variables quantitatives:

- La statistique la plus utilisée dans le contexte de deux séries numériques est la corrélation. Pour la définir, la notion de covariance doit être introduite.
- On appelle covariance des séries numériques $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$ la valeur

$$\begin{aligned} s_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}, \\ &= \frac{\sum_{i=1}^{n} x_i y_i}{n} - (\overline{x} \cdot \overline{y}). \end{aligned}$$

- Le coefficient corrélation (linéaire) des séries numériques $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$ est défini par

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Interprétation du coefficient de corrélation

- On a: $-1 \leq \rho_{xy} \leq 1$.
- Si $\rho_{xy} = +1$ alors il y a corrélation parfaite (positive) entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont alignés sur une droite de pente positive.
- Si $\rho_{xy} = -1$ alors il y a corrélation parfaite (négative) entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont alignés sur une droite de pente négative.
- Si $\rho_{xy} = 0$ alors il n'y a pas de corrélation entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont distribués "au hasard" dans le plan.

3.3. La droite de régression linéaire

Le but de la régression simple (resp. multiple) est d'expliquer une variable Y à l'aide d'une variable X (resp. plusieurs variables $X_1, \cdots, X_q$). La variable Y est appelée variable dépendante, ou variable à expliquer et les variables X_j, $j = 1, \cdots, q$ sont appelées variables indépendantes, ou variables explicatives.

Remarques:

- Il est indispensable de commencer par l'étude corrélation entre X et Y avant de chercher une ligne de régression entre X et Y.
- La régression diffère de l'analyse de la corrélation où toutes les variables jouent un rôle symétrique (pas de variable dépendante versus indépendante).
- Toutefois, tout comme dans le contexte de l'analyse de la corrélation, il faut être prudent lorsqu'on formule des relations de causalité! L'existence d'une relation entre X et Y n'implique pas nécessairement une relation de causalité entre elles.

Un modèle de régression simple (resp. multiple) est défini par l'équation:

$$Y = f(X) + \text{Erreur (resp. } Y = f(X_1, \cdots, X_q) + \text{Erreur)}.$$

- Si $f(x) = ax + b$, on parle de la régression **linéaire** simple.
- Les paramètres a et b sont inconnus et estimés par la méthode des moindres carrés.

$$\begin{aligned} \widehat{a} &= \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}, \\ \widehat{b} &= \overline{Y} - \widehat{a}\overline{X}, \end{aligned}$$

où $\overline{X}$ et $\overline{Y}$ désignent les moyennes de X et Y, respectivement. Le point $M\left(\overline{X}, \overline{Y}\right)$ est appelé point moyen.

- La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée la droite de régression linéaire .
- Le couple $\left(\widehat{a}, \widehat{b}\right)$ minimise la somme des carrés des résidus e_i:

$$\sum_{i=1}^{n} e_i^2 \quad = \quad \sum_{i=1}^{n} (Y_i - aX_i - b)^2, \quad a, b \in \mathbb{R}.$$

- Un des buts de la régression est de proposer des **prédictions** pour la variable à expliquer Y lorsque nous avons de nouvelles valeurs de X.
- Pour une nouvelle valeur x_0, on prévoit la valeur y_0 associée

$$y_0 = \widehat{a}x_0 + \widehat{b}.$$

Exemple 1: Le tableau suivant donne l'évolution du nombre de spectateurs (en millions) dans les salles de cinéma en France sur une période de 7 ans.

Année	1989	1993	1994	1995	1996
Rang (X_i)	0	4	5	6	7
Nombre de spectateurs (Y_i)	120,9	132,7	124,5	130,2	136,3

- Représenter la série statistique (X_i, Y_i) par un nuage de points, et préciser le point moyen M.
- Donner une équation de la droite de régression (Δ) de Y en X par la méthode des moindres carrés.
- Si l'évolution constatée s'est poursuivie jusqu'à la fin du XX^e siècle, donner une estimation du nombre de spectateurs dans les salles de cinéma en France en l'an 2000.

Exemple 2: On a relevé pour chacune des années t de 1920 à 1929, numérotées de 1 à 10, la température moyenne X des mois d'été (en degrés centigrades) et la mortalité infantile Y (nombre de décès d'enfants de moins d'un an pour 1000 naissances vivantes).

t	1	2	3	4	5	6	7	8	9	10
X	15,9	18,8	15,4	18	14,6	16,2	17,9	16,5	18,1	19,8
Y	98	116	87	96	85	89	97	83	91	95

- Étudier la corrélation entre X et Y.
- Déterminer la droite de régression linéaire de Y en X.

4. Techniques de prévision et estimation

4.1. Méthodes quantitatives de prévision
4.2. Rappel: Probabilités
4.3. Estimation des paramètres

- L'étude d'une série chronologique permet d'**analyser**, de **décrire** et d'**expliquer** un phénomène au cours du temps et d'en tirer des conséquences pour des prises de décision (marketing,...etc).
- Cette étude permet aussi de faire un **contrôle**, par exemple pour la gestion des stocks, le contrôle d'un processus chimique...
- L'un des objectifs principaux de l'étude d'une série chronologique est la **prévision** qui consiste à prévoir les valeurs futures (X_{t+h}, $h = 1, 2, 3, \cdots$) de la série chronologique à partir de ses valeurs observées jusqu'au temps t : $X_1, X_2, \cdots, X_t$.

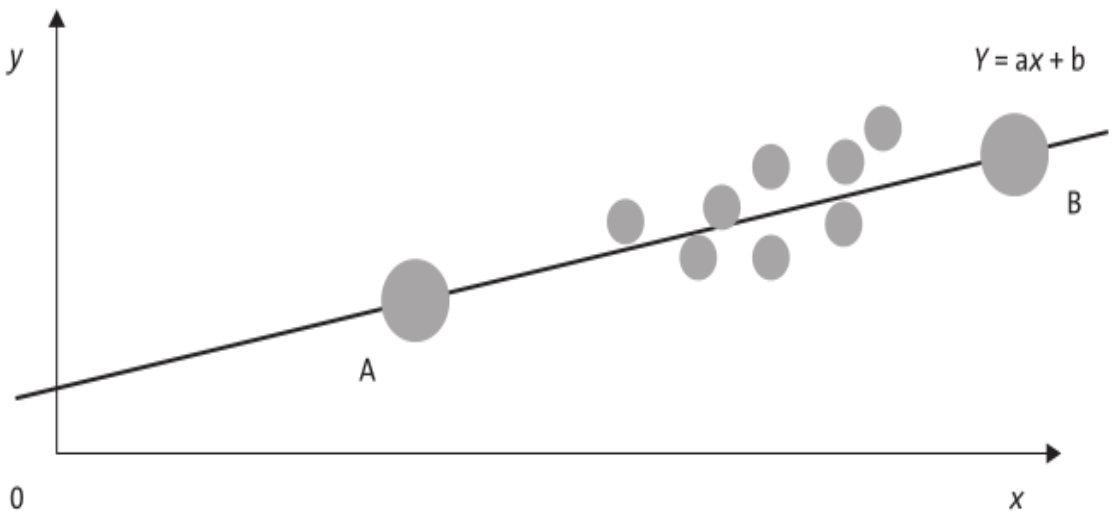
- La prédiction de la série chronologique au temps $t + h$ est notée $\widehat{X}_t(h)$. En général, elle est différente de la valeur réelle X_{t+h} que prend la série au temps $t + h$.
- Pour mesurer cette différence, on définira l'erreur de prédiction par la différence $\widehat{X}_t(h) - X_{t+h}$.
- La qualité de la prévision dépend de la façon dont évolue la série. Plus la série est fonction "régulière "du temps, plus il sera facile de prévoir.

- En économie, les données constituent souvent des séries d'observations sur une ou plusieurs variables faites à différentes dates.
- Les observations ne sont pas indépendantes, et constituent une série de données qui se suit dans un ordre chronologique, sous une forme remarquablement significative.
- On appelle une **série chronologique** (on dit aussi chronique ou série temporelle) toute suite d'observations X_t $t = 1, 2, \cdots, n$; le nombre n est appelé la longueur de la série. L'indice temps peut être selon les cas l'heure, le jour, le mois, l'année, etc...

4.1. Méthodes quantitatives de prévision: ajustements linéaires

4.1.1. L'ajustement par La méthode des points extrêmes

- La méthode des points extrêmes est une méthode d'ajustement linéaire d'équation $y = ax + b$ déterminée à partir des coordonnées des deux points extrêmes d'une série d'observations sur la période analysée.



Exemple 1: On considère la série chronologique suivante.

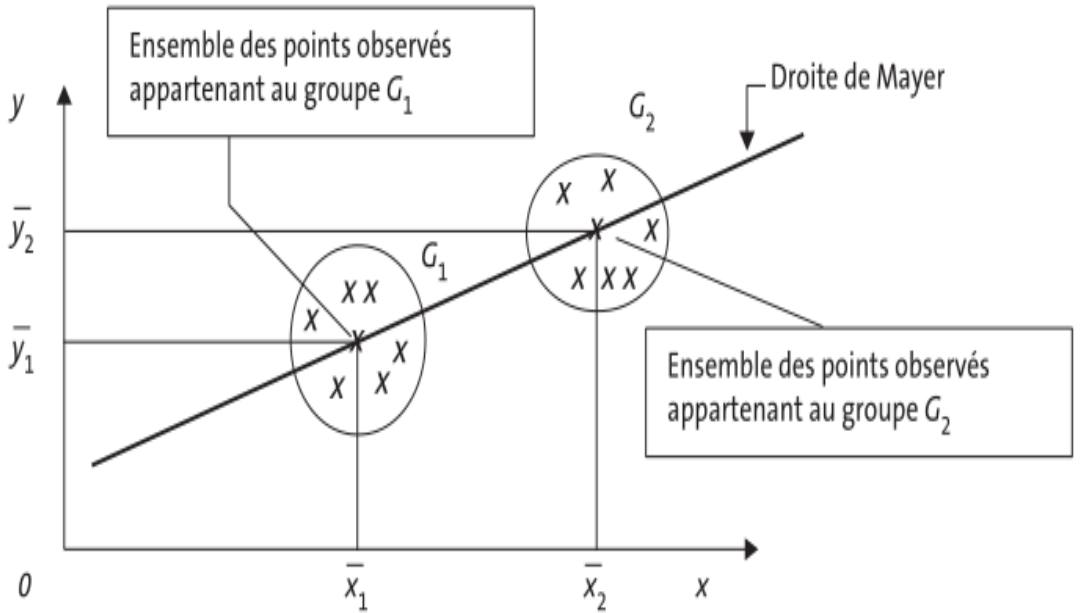
Mois t	1	2	3	4	5	6	7
Ventes (milliers de MAD) y_t	120	155	125	202	180	235	240

1. Représenter graphiquement la série par un nuage de points.
2. Déterminer l'équation de la droite de tendance par la méthode des points extrêmes.
3. Représenter cette équation sur le même graphique.
4. Déterminer la prévision des ventes pour le mois 8.

4.1.2. L'ajustement par La méthode des points moyens (méthode de Mayer)

Cette méthode consiste à:

- Partager la série statistique en deux groupes G_1, G_2.
- Calculer les coordonnées des points moyens M_1, M_2 de chaque groupe.
- Déterminer la droite de tendance passant par M_1 et M_2.



Il est important de noter que la méthode des points extrêmes, ainsi que celle de Mayer, sont peu précises et ne sont pertinentes qu'en présence d'une très grande stabilité des observations.

4.1.3. L'ajustement par La méthode des moindres carrés

- La méthode des moindres carrés a pour objectif d'ajuster les données statistiques par une droite de la forme $y = ax + b$.
- Graphiquement, la droite d'ajustement des moindres carrés cherche à minimiser la somme des carrés des distances entre la valeur observée et la valeur ajustée: $\sum_i (y_i - ax_i - b)^2$.
- Les paramètres a et b sont donnés par:

$$\begin{aligned} a &= \frac{s_{xy}}{s_x^2}, \\ b &= \overline{y} - a\overline{x}. \end{aligned}$$

- **La méthode des moindres carrés est considérée comme étant la plus fiable car elle minimise la somme des carrés des distances entre la valeur observée et la valeur ajustée.**

4.1.4. Les moyennes mobiles

- La méthode des moyennes mobiles est une technique de lissage des données. Son principe est de substituer une série de valeurs observées par leur moyenne.
- Cette moyenne est calculée en prenant par exemple, trois valeurs (nous dirons qu'il s'agit de moyennes mobiles d'ordre 3), quatre valeurs (moyennes mobiles d'ordre 4), etc.
- La périodicité dépend de la saisonnalité du chiffre d'affaires. Si la périodicité est donnée en trimestre (périodicité d'ordre 3), on calcule les moyennes sur les **trois trimestres consécutifs** et en les attribuant au 2 ème trimestre :

$$\frac{y_1 + y_2 + y_3}{3}.$$

- La prévision par **moyenne mobile** d'ordre k consiste à calculer la moyenne des k dernières données, et à l'employer comme prévision.

$$\widehat{y}_{t+1} = \frac{\sum_{j=0}^{k-1} y_{t-j}}{k}.$$

4.1.5. Les moyennes mobiles pondérées

- Avec la méthode de la moyenne mobile pondérée, on applique un coefficient de pondération à chaque terme de la série chronologique prise en compte dans la moyenne.
- La somme des coefficients doit être égale à 1.
- La prévision par **moyenne mobile pondérée** d'ordre k est donnée par

$$\widehat{y}_{t+1} = \sum_{j=0}^{k-1} \alpha_j y_{t-j}, \quad \sum_{j=0}^{k-1} \alpha_j = 1.$$

- Cette méthode a pour avantage de donner plus de poids aux données récentes.

4.1.6. Lissage exponentiel simple

- On dispose de $y_1, \cdots, y_n$, $n \geq 1$ et on veut estimer y_{n+1}. Pour $\alpha \in]0;1[$, on définit la prévision par **lissage exponentiel simple**

$$\widehat{y}_{n+1} = \alpha \sum_{j=0}^{n-1} (1-\alpha)^j y_{n-j}.$$

- Plus α est petit, plus on donne d'importance aux observations anciennes.
- La somme des poids ne fait pas 1.
- On peut calculer par récurrence à l'aide de la formule de mise à jour :

$$\widehat{y}_{n+1} = \alpha y_n + (1-\alpha)\widehat{y}_n.$$

- Pour choisir α, on peut se servir de la remarque précédente ou calculer, pour tout α, les estimateurs calculés avec le paramètre α : $\widehat{y}_{t+1}(\alpha)$. On regarde ensuite l'erreur quadratique

$$E_2(\alpha) \quad = \quad \sum_{t=1}^{n-1} \left(\widehat{y}_{t+1}(\alpha) - y_{t+1}\right)^2 .$$

- Si cette erreur est petite, c'est que le paramètre α produit des prédictions performantes, au vu des données $y_1, \cdots, y_n$. On peut choisir entre plusieurs paramètres $\alpha_1, \cdots, \alpha_p$ en prenant

$$\alpha \quad = \quad \arg\min_{\alpha_i} E_2(\alpha_i).$$

4.2. Rappel: Probabilités

- Une **expérience** est dite **aléatoire** si on ne peut pas prédire a priori son résultat. On note ω un résultat possible de cette expérience aléatoire. L'ensemble de tous les résultats possibles est appelé **univers** et noté Ω.
- On appelle événement tout sous-ensemble de l'univers Ω.
- L'ensemble des événements est noté $\mathcal{F}$.

Terminologie des événements :

- $\emptyset$ est appelé **événement impossible**.
- Ω est appelé **événement certain**.
- Tout singleton $\{\omega\}$, où $\omega \in \Omega$, est appelé **événement élémentaire**.
- Nous appelons **événements incompatibles** ou **disjoints** deux événements A et B tels que $A \cap B = \emptyset$, c'est-à-dire qu'il est impossible que A et B se réalisent simultanément.

Proposer des exemples d'expériences aléatoires, et préciser l'univers Ω.

Dorénavant, On suppose que l'ensemble des événements est noté $\mathcal{F}$ vérifie les conditions suivantes:

- $\emptyset \in \mathcal{F}$.
- Si $A \in \mathcal{F}$, alors son complémentaire $A^c \in \mathcal{F}$ (où A^c est le complémentaire de A dans Ω).
- Toute réunion dénombrable d'événements de $\mathcal{F}$ est aussi un événement de $\mathcal{F}$.

Dans ce cas le couple $(\Omega, \mathcal{F})$ est appelé **espace probabilisable**.

Probabilité

On appelle probabilité sur $(\Omega, \mathcal{F})$ une application $\mathbb{P}$ de $\mathcal{F}$ dans $[0,1]$ vérifiant les deux propriétés suivantes:

- $\mathbb{P}(\Omega) = 1$.
- Pour toute suite $(A_i)_{i\in I}$ dénombrable d'événements de $\mathcal{F}$ deux à deux incompatibles, on a

$$\mathbb{P}\left(\cup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i).$$

Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé **espace probabilisé**.

En théorie des probabilités, le terme **modéliser** désigne l'opération qui consiste à associer à une expérience aléatoire le triplet $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \mathbb{P})$.

Probabilité sur un univers fini:
On considère une expérience aléatoire modélisée par $(\Omega, \mathcal{F}, \mathbb{P})$ avec $\Omega = \{\omega_1, \cdots, \omega_n\}$. Pour tous événements A et B de $\mathcal{F}$, on a

- $0 \leq \mathbb{P}(A) \leq 1$.
- Si $A \subset B$, alors $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- Si A et B sont incompatibles, alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
- En général, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Équiprobabilité et probabilité uniforme
Nous disons qu'il y a équiprobabilité lorsque les probabilités de tous les événements élémentaires sont égales. Dans ce cas, P est la probabilité uniforme sur $(\Omega, \mathcal{F})$.

Conséquence: S'il y a équiprobabilité, pour tout événement A, nous avons alors

$$\mathbb{P}(A) \quad = \quad \frac{Card(A)}{Card(\Omega)} \quad = \quad \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}.$$

Variable aléatoire réelle

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. Nous appelons variable aléatoire réelle (v.a.r.) toute application de Ω dans $\mathbb{R}$ telle que : pour tout intervalle I de $\mathbb{R}$, l'ensemble

$$X^{-1}(I) \;=\; \{\omega \in \Omega \;:\; X(\omega) \in I\} \in \mathcal{F}.$$

Notations

- Une variable aléatoire est notée avec une lettre **majuscule** et sa réalisation (quantité déterministe) avec une lettre **minuscule**.
- On note $X(\Omega)$ l'ensemble des valeurs prises par la variable aléatoire X définie sur l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.
- La variable X est dite discrète (resp. continue) si $X(\Omega)$ est discret (resp. continu).

- L'ensemble $X^{-1}(\{x\})$ se note $\{X = x\}$.
- L'ensemble $X^{-1}(]-\infty, a])$ se note $\{X \leq a\}$.
- L'ensemble $X^{-1}(]a, b])$ se note $\{a < X \leq b\}$.

Loi d'une variable aléatoire discrète, espérance et variance
Soit X une variable aléatoire réelle définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. Nous appelons loi de la variable aléatoire X la donnée d'une suite numérique $(P_X(k))_{k\in X(\Omega)}$ telle que

- $\mathbb{P}(X = k) = P_X(k)$ et $0 \leq P_X(k) \leq 1$.
- $\sum_{k\in X(\Omega)} P_X(k) = 1$.
- $\mathbb{P}(X \leq x) = \sum_{k\leq x} P_X(k)$.

L'espérance et **la variance** de X sont définies respectivement par

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k\in X(\Omega)} kP_X(k), \\ \mathbb{V}ar(X) &= \mathbb{E}\left(X - \mathbb{E}(X)\right)^2. \end{aligned}$$

Indépendance de deux variables aléatoires discrètes
Deux v.a.d. X et Y sont dites indépendantes si:

$$\mathbb{P}(X = x, Y = y) \quad = \quad \mathbb{P}(X = x)\mathbb{P}(Y = y), \quad \forall x \in X(\Omega), \ \forall y \in Y(\Omega).$$

L'indépendance de X et Y signifie qu'il n'y a aucune influence l'un sur l'autre.

Exercice 1.

Soient X et Y deux variables aléatoires discrètes définies sur un même univers Ω, et $a, b \in \mathbb{R}$. On pose $X(\Omega) = \{x_1, \cdots, x_n\}$ et $Y(\Omega) = \{y_1, \cdots, y_m\}$. Vérifier que

(1) $\mathbb{E}(aX + Y + b) = a\mathbb{E}(X) + \mathbb{E}(Y) + b$.

(2) Si X et Y sont deux variables indépendantes, alors

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

(3) $\mathbb{V}ar(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$.

(4) Si X et Y sont deux variables indépendantes, alors

$$\mathbb{V}ar(aX + Y + b) = a^2\mathbb{V}ar(X) + \mathbb{V}ar(Y).$$

4.3. Estimation des paramètres

Échantillon aléatoire

Un échantillon aléatoire est une suite de variables aléatoires $X_1, \cdots, X_n$ indépendantes et de même loi qu'un caractère X d'une population.

Paramètre

- Un paramètre est un nombre qui décrit une caractéristique de la population étudiée.
- Citons, à titre d'exemples, **la moyenne** μ, **la variance** σ^2, **la médiane** M et **la proportion** p d'une population.
- Notons que les paramètres sont souvent inconnus.

Estimateur et qualité d'un estimateur

- Un **estimateur** est une fonction $T(X_1, \cdots, X_n)$ de l'échantillon qui permet d'estimer un paramètre θ de la population.
- Soient $x_1, \cdots, x_n$ des observations obtenues à partir d'un échantillon. La valeur $T(x_1, \cdots, x_n)$ est appelée **estimation ponctuelle** du paramètre θ.

Exemple 1.

On considère une population partiellement observée de moyenne μ et de variance σ^2. Soit p une proportion d'une caractéristique qualitative dans cette population. On peut estimer ces paramètres à partir des échantillons de cette population par

Paramètre	μ	σ^2	p
Estimateur	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.	$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$, $S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$.	$\widehat{p} = \frac{S_n}{n}$, où S_n désigne le nombre d'individus de l'échantillon qui possèdent la caractéristique d'intérêt.

- Un estimateur $T = T(X_1, \cdots, X_n)$ d'un paramètre θ est dit **sans biais** si $\mathbb{E}(T) = \theta$.
- Si $\mathbb{E}(T) \longrightarrow \theta$, lorsque $n \to +\infty$, on dit que T est asymptotiquement sans biais.
- Notons qu'un estimateur sans biais ne surestime ni sous-estime systématiquement le paramètre. On dit d'un estimateur sans biais qu'il est bien centré.

Exercice 2.

Vérifier que

(a) Les estimateurs $\overline{X}$, S_X^2 et $\widehat{p}$ sont sans biais.

(b) L'estimateur s_X^2 est asymptotiquement sans biais.

Efficacité d'un estimateur

Soient T_1 et T_2 deux estimateurs sans biais d'un paramètre inconnu θ. On dit que T_1 est plus **efficace** que T_2 si $\mathbb{V}ar(T_1) \leq \mathbb{V}ar(T_2)$.

Exercice 3.

L'écart quadratique moyen d'un estimateur T du paramètre θ est défini par $EQM = \mathbb{E}\left[(T-\theta)^2\right]$, et son biais est défini par $b(T) = \mathbb{E}(T) - \theta$. Vérifier que

1. $EQM = \mathbb{V}ar(T) + b(T)^2$.
2. Soient T_1 et T_2 deux estimateurs sans biais du paramètre θ. On pose $T' = \alpha T_1 + (1-\alpha) T_2$, avec $\alpha \in [0,1]$.
   - **(a)** Calculer l'écart quadratique moyen de chacun des estimateurs T_1, T_2 et T'.
   - **(b)** Trouver la valeur de α pour que T' soit plus efficace que T_1 et T_2.

Estimation par intervalle de confiance

Un intervalle de confiance de seuil (ou niveau de signification) $\alpha \in]0,1[$ pour un paramètre θ est un intervalle aléatoire I tel que $\mathbb{P}(\theta \in I) \geq 1 - \alpha$.

- On cherche souvent un intervalle de confiance de la forme $\left[\widehat{\theta} - \varepsilon, \widehat{\theta} + \varepsilon\right]$, où $\widehat{\theta}$ est un estimateur du paramètre θ.
- La quantité $1 - \alpha$ est appelée niveau de confiance, et la marge d'erreur ε est appelée précision de l'estimateur $\widehat{\theta}$.

Dans la suite on fixe α et $z_{1-\alpha/2}$

Seuil α	$0,05$	$0,01$
$z_{1-\alpha/2}$	$1,96$	$2,575$

Dans une population de moyenne μ, de variance σ^2 et de proportion p, on considère les estimateurs $\overline{X}$, S_X^2 et $\widehat{p}$ construits à partir des échantillons de taille $n \geq 30$. Les intervalles de confiance des paramètres μ et p sont

Paramètre	**Population de taille infinie**	**Population de taille finie N**
$\boldsymbol{\mu}$ σ^2 est connue:	$\textbf{IC} = \overline{X} \pm z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	$\textbf{IC} = \overline{X} \pm z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$
$\boldsymbol{\mu}$ σ^2 est inconnue:	$\textbf{IC} = \overline{X} \pm z_{1-\alpha/2}\dfrac{S_X}{\sqrt{n}}$	$\textbf{IC} = \overline{X} \pm z_{1-\alpha/2}\dfrac{S_X}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$
$\boldsymbol{p}$ $np \geq 5$ et $n(1-p) \geq 5$	$\textbf{IC} = \widehat{p} \pm z_{1-\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$	$\textbf{IC} = \widehat{p} \pm z_{1-\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{\dfrac{N-n}{N-1}}$

Choix de la taille d'échantillon

- La qualité d'un intervalle de confiance se mesure par son niveau de confiance $1 - \alpha$ et sa marge d'erreur ε.
- Un choix adéquat de la taille de l'échantillon permet de contrôler simultanément ces deux facteurs.