Modélisation au Lycée en Statistique et Probabilités

Jeanne Fine
Professeure de Mathématiques
I.U.F.M. de Midi-Pyrénées
Laboratoire de Statistique et Probabilités,
Université Toulouse 3

L’enseignement des sciences au lycée
(BO Hors Série n°6 du 12 août 1999)

« Le développement des sciences se fait par un va-et-vient entre :
- l’observation et l’expérience d’un côté,
- la conceptualisation et la modélisation de l’autre, …

…
L’exercice de modélisation du réel est sans doute l'étape la plus importante et la plus difficile de la démarche scientifique.
…
Les mathématiques se rapprochent des sciences expérimentales grâce à l'expérimentation numérique, à la simulation, et à ce que l'on peut appeler la démonstration empirique. »

La Statistique : connaissance et maîtrise de l'aléatoire
(Conférence Maths2000, IUFM Midi-Pyrénées, 25 janvier 2000)

Quatre parties :

1. la Statistique Descriptive
synthétiser l'information recueillie

2. le Calcul des Probabilités
étudier les phénomènes aléatoires

3. la Statistique Inférentielle
inférer à la population les résultats observés
sur un échantillon aléatoire

4. la Modélisation Aléatoire
décrire, expliquer et prévoir

Modélisation au Lycée
en Statistique et Probabilités

Exemples en Statistique :

1. Modèle de régression linéaire simple
2. Modèles d'analyse de séries chronologiques

Exemple en Probabilités :

3. Modèle probabilisé associé à une expérience aléatoire et simulation

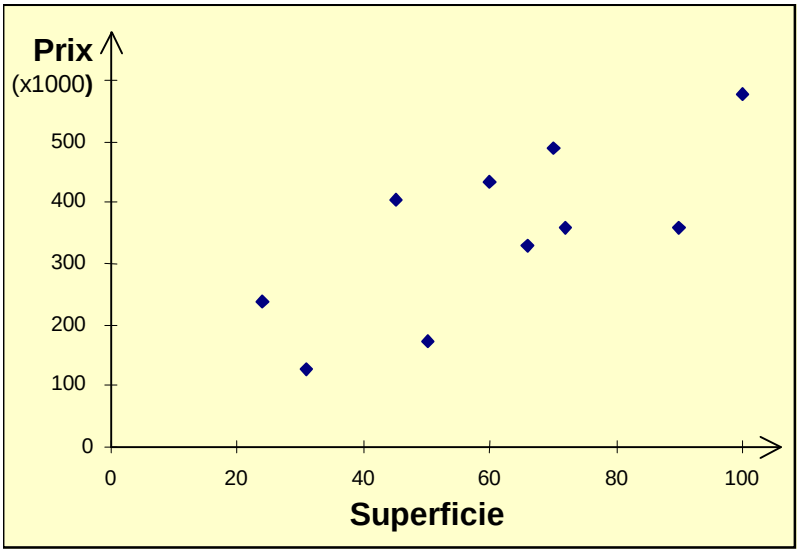
1. Modèle de régression linéaire simple

1.1. Les données

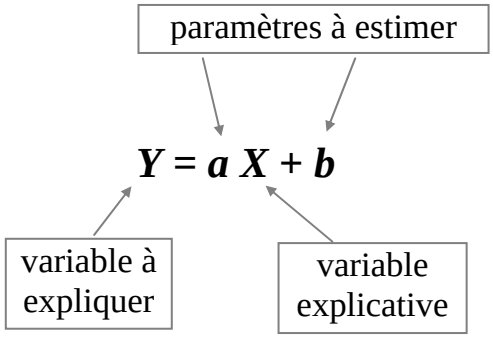
Population : les "ventes" d'appartements à Toulouse en 1996
Le prix en francs : Y
La superficie en mètres carrés : X

Observations : $(X_i, Y_i) \quad i = 1, \ldots, n$

Prix Y (en F)	Superficie X (en m^2)
239 500	24
126 500	31
405 000	45
173 000	50
435 000	60
330 000	66
360 000	72
490 000	70
360 000	90
576 000	100



1.2. Le modèle



Introduction des erreurs

$$Y_i = a X_i + b + E_i \quad i = 1, \ldots, n$$

1.3. Estimation des paramètres par la méthode des moindres carrés des erreurs

$$Y_i = aX_i + b + E_i \quad i = 1,\ldots,n$$

Les valeurs de *a* et de *b* qui rendent minimum la moyenne des carrés des erreurs :

$$f(a,b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - aX_i - b)^2$$

sont : $\hat{a} = \dfrac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$ et $\hat{b} = \bar{Y} - \hat{a}\bar{X}$

La droite d'équation $y = \hat{a}x + \hat{b}$ est appelée droite de régression linéaire de *Y* en *X*.

Ajustement du modèle
et interprétation

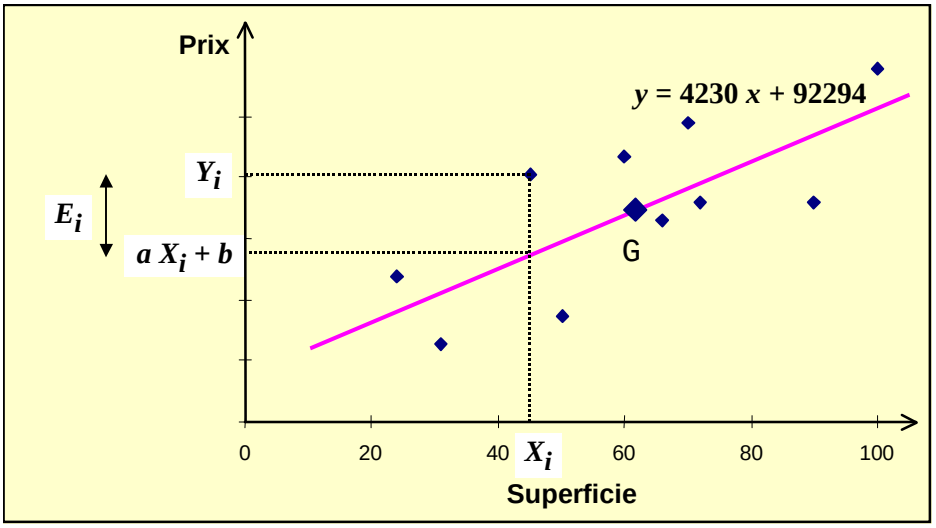
On obtient ici :

$$\overline{X} = 61 \qquad \overline{Y} = 349\,500$$

$$y = 4230\,x + 92\,294$$

Le prix moyen des appartements est de 349 500 F pour une superficie moyenne de 61 m^2 ;

- pour chaque m^2 en plus, le prix augmente en moyenne de 4230 F
- pour chaque m^2 en moins, le prix diminue en moyenne de 4230 F

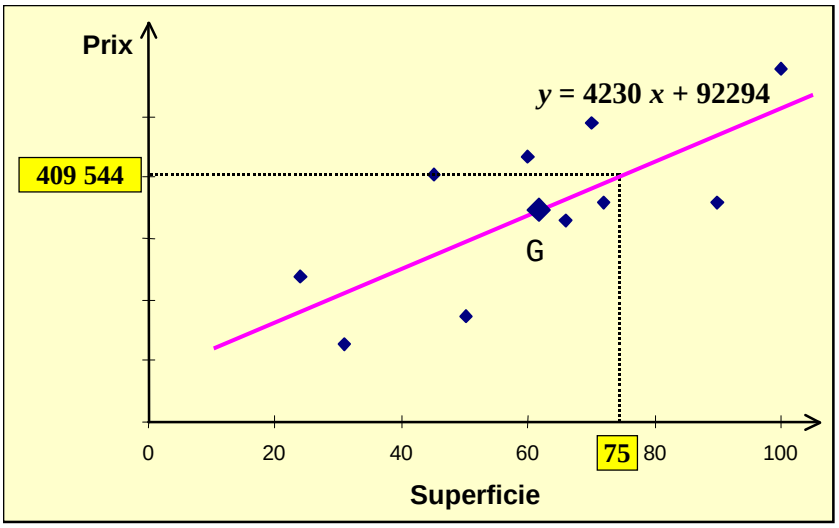


1.4. Prévision

Quel prix prévoir pour un appartement de 75 m^2 ?

Le modèle ajusté est : $\hat{Y} = 4230\,X + 92\,294$

Pour X = 75 on a : $\hat{Y} = 4230 \times 75 + 92\,294 = 409\,544$ F



1.5. Décomposition de la variance

On construit les deux variables :

$\hat{Y} = \hat{a}\,X + \hat{b}$ (le modèle) et

$\hat{E} = Y - \hat{Y}$ (l'erreur d'ajustement).

	X (Surface)	Y (Prix)	$\hat{Y} = \hat{a}\,X + \hat{b}$	$\hat{E} = Y - \hat{Y}$
	24	239 500	193 823	45 677
	31	126 500	223 435	-96 935
	45	405 000	282 660	122 340
	50	173 000	303 812	-130 812
	60	435 000	346 116	88 884
	66	330 000	371 498	-41 498
	72	360 000	396 880	-36 880
	70	490 000	388 419	101 581
	90	360 000	473 027	-113 027
	100	576 000	515 330	60 670
Moyenne	61	349 500	349 500	0
Variance	523.56	17 501 500 000	9 369 624 347	8 131 875 654

On vérifie $\overline{\hat{Y}} = \overline{Y}$, $\overline{\hat{E}} = 0$, $\mathrm{cov}\left(\hat{Y}, \hat{E}\right) = 0$

et $$\boxed{\mathbf{var}\left(\boldsymbol{Y}\right) = \mathbf{var}\left(\hat{\boldsymbol{Y}}\right) + \mathbf{var}\left(\hat{\boldsymbol{E}}\right)}$$

$R^2 = \dfrac{\mathrm{var}\left(\hat{Y}\right)}{\mathrm{var}\left(Y\right)}$ (= 0.535) est la part de variance expliquée par le modèle ;

$1 - R^2$ (= 0.465) est la part résiduelle.

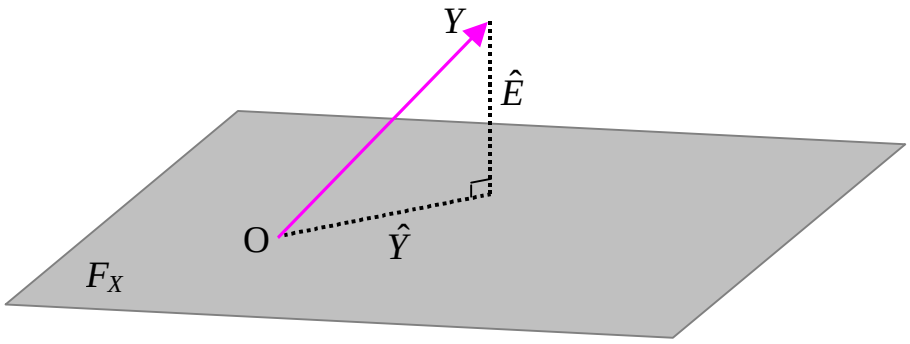
1.6. Interprétation géométrique

Dans $\mathbb{R}^n$ muni du produit-scalaire : $\langle x,y\rangle_n = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i$

$$f(a,b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - a\,X_i - b)^2 = \|Y - a\,X - b\mathbb{1}_n\|_n^2$$

avec $Y = (Y_1,\ldots,Y_n)$, $X = (X_1,\ldots,X_n)$ et $\mathbb{1}_n = (1,\ldots,1)$.

Soit $F_X = \left\{ a\,X + b\mathbb{1}_n \,\middle|\, (a,b) \in \mathbb{R}^2 \right\}$ le plan engendré par X et $\mathbb{1}_n$.



La recherche de a et b minimisant $f(a,b)$ est la recherche du vecteur de F_X le plus proche de Y.
La solution $\hat{Y} = \hat{a}\,X + \hat{b}\mathbb{1}_n$ est la projection orthogonale de Y sur F_X .

1.7. Généralisation

Plusieurs variables explicatives réelles $\left(X_1,\ldots,X_p\right)$ ⟹ Modèle de régression linéaire multiple

Une variable explicative catégorielle à p modalités ⟹ Analyse de variance à un facteur

$\left(X_1,\ldots,X_p\right)$ les indicatrices des modalités

Plusieurs variables explicatives réelles ou catégorielles ⟹ Analyse de covariance

1.8. Coefficient de corrélation linéaire

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma(X)\,\sigma(Y)} \qquad -1 \le \rho \le 1$$

$$|\rho| = 1 \quad \Leftrightarrow \quad \exists\, a \in \mathbb{R}_*^+ \text{ et } b \in \mathbb{R} \quad Y = aX + b$$

On a ici : $\rho = 0.73 \quad \left(= \sqrt{R^2}\right)$

1.9. Origine historique du mot "régression"

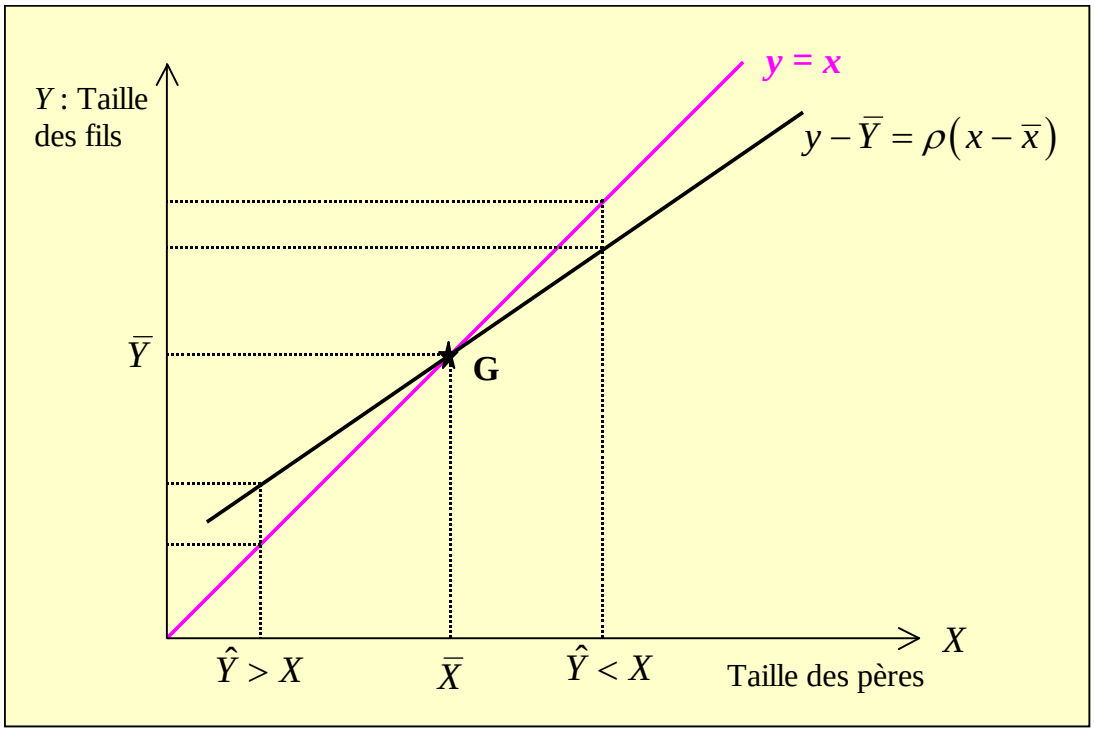
$$Y - \bar{Y} = \frac{\text{cov}(X,Y)}{\text{var}(X)}\left(X - \bar{X}\right) + E$$

Galton (1822, 1911) étudie la taille des fils (Y) en fonction de la taille des pères (X).

Bien que les moyennes (resp. les variances) des tailles des fils et des pères soient égales, il observe une **régression vers la moyenne**.

En effet, l'équation de la droite de régression linéaire de Y en X s'écrit alors :

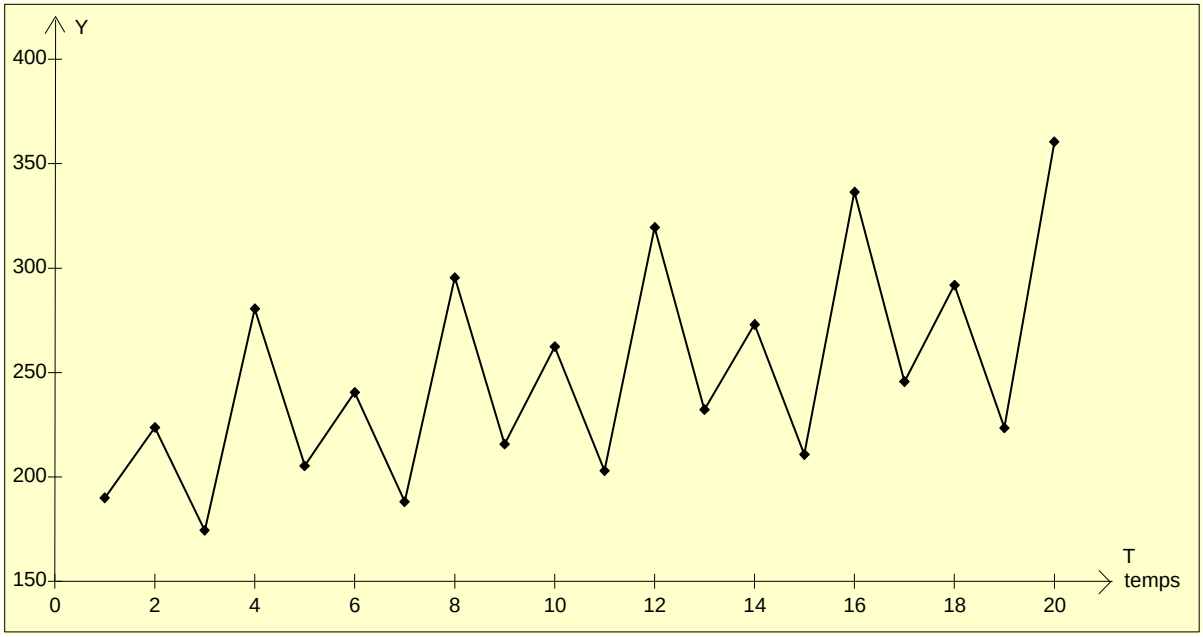
$$y - \bar{Y} = \rho\left(x - \bar{X}\right) \text{ et } \rho < 1$$



2. Modèles d'analyse de séries chronologiques

2.1. Les données

Représentation graphique du chiffre d'affaire Y d'une entreprise en milliers de francs pour chaque trimestre de cinq années consécutives Y(t), t = 1, … , 20.



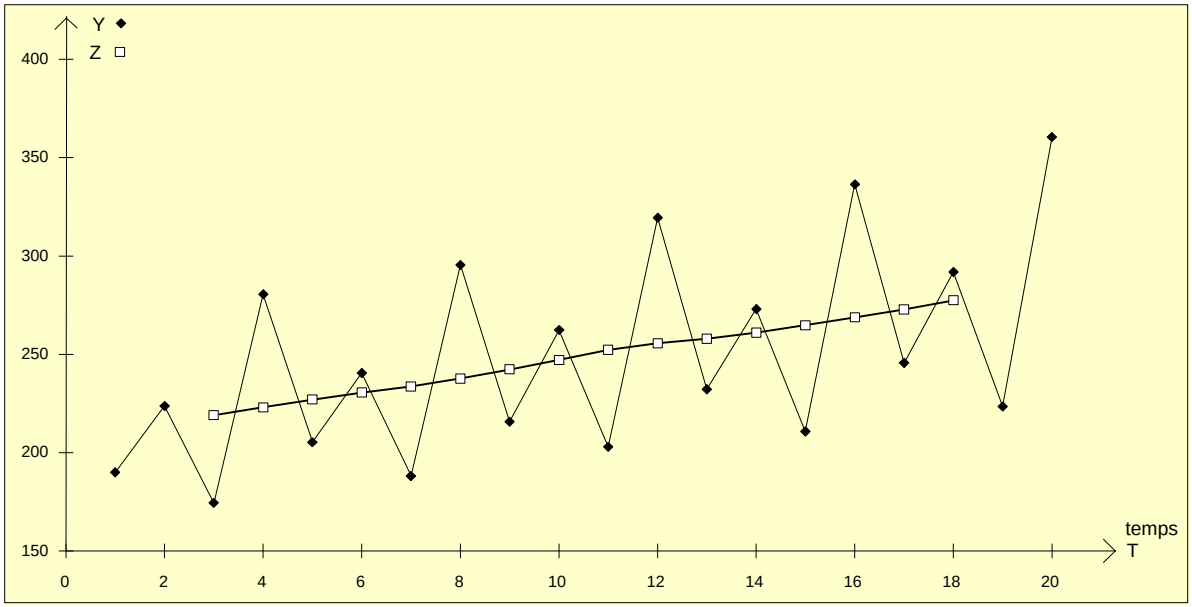
2.2. Les deux composantes de la série *Y* (*t*)

- une composante tendancielle (ou « tendance » ou « trend »), *X* (*t*).
- une composante saisonnière (ici trimestrielle), *S*(*t*).

2.3. Lissage de la série : série des moyennes mobiles

La moyenne mobile d'ordre 4 est définie par :

$$Z(t)=\frac{1}{4}\left[\frac{1}{2}Y(t-2)+Y(t-1)+Y(t)+Y(t+1)+\frac{1}{2}Y(t+2)\right] \quad t=3,...,18$$



2.4. Détermination de la tendance

On décide d'une tendance linéaire.

On choisit pour tendance la droite de régression linéaire de Y en T :

$$X(t)=4.68\,t+199.52 \quad t=1,\ldots,20$$

2.5. Détermination de la composante saisonnière

La composante saisonnière est ici trimestrielle, périodique de période 4

Modèle additif :

$$Y(t) = X(t) + S_a(t),\ t = 1,\ldots,20$$

On calcule les différences

$$Y(t) - X(t)$$

Modèle multiplicatif :

$$Y(t) = X(t) \times S_m(t),\ t = 1,\ldots,20$$

On calcule les quotients

$$Y(t) / X(t)$$

Pour chacun des deux modèles :
le coefficient d'un trimestre est la moyenne des 5 valeurs obtenues pour ce trimestre.

2.6. Résultat de l'ajustement des modèles et erreurs d'ajustement

Modèle additif :

$$\hat{Y}_a(t) = 4.68\,t + 199.52 + \begin{vmatrix} -3.8 & 1°\text{tr} \\ 12.0 & 2°\text{tr} \\ -51.0 & 3°\text{tr} \\ 62.8 & 4°\text{tr} \end{vmatrix}$$

Modèle multiplicatif :

$$\hat{Y}_m(t) = (4.68\,t + 199.52) \times \begin{vmatrix} 0.90 & 1°\text{tr} \\ 1.05 & 2°\text{tr} \\ 0.80 & 3°\text{tr} \\ 1.25 & 4°\text{tr} \end{vmatrix}$$

Les erreurs d'ajustement

Modèle additif

$$E_a(t) = Y(t) - \hat{Y}_a(t)$$
$$t = 1, \ldots, 20$$

Modèle multiplicatif

$$E_m(t) = Y(t) - \hat{Y}_m(t)$$
$$t = 1, \ldots, 20$$

2.7. Prévisions

Les prévisions pour la 6ème année

6ème année	**Modèle additif** $\hat{Y}_a(t)$	**Modèle multiplicatif** $\hat{Y}_m(t)$
1er trim.	274.0	268.0
2ème trim.	314.5	317.6
3ème trim.	256.2	245.7
4ème trim.	374.6	389.8

La réalité

6ème année	$Y(t)$
1er trim.	272.1
2ème trim.	315.3
3ème trim.	250.2
4ème trim.	381.0

2.8. Les erreurs de prévision

Modèle additif

$$E_a(t) = Y(t) - \hat{Y}_a(t)$$
$$t = 21,\ldots,24$$

Modèle multiplicatif

$$E_m(t) = Y(t) - \hat{Y}_m(t)$$
$$t = 21,\ldots,24$$

6ème année	Modèle additif	Modèle multiplicatif
1er trim	-1.9	4.1
2ème trim	0.8	-2.3
3ème trim	-6.0	4.5
4ème trim	6.4	-8.8

2.9. Critères de qualité d'un modèle

- Critère des moindres carrés des ***erreurs d'ajustement***,

$$\sum_{t=1}^{20}\left(\hat{E}_a(t)\right)^2 = 800 \qquad \sum_{t=1}^{20}\left(\hat{E}_m(t)\right)^2 = 431$$

$\rightarrow$ le modèle multiplicatif est "meilleur".

- Critère des moindres carrés des ***erreurs de prévision***

$$\sum_{t=21}^{24}\left(\hat{E}_m(t)\right)^2 = 81 \qquad \sum_{t=21}^{24}\left(\hat{E}_m(t)\right)^2 = 120$$

$\rightarrow$ le modèle additif est "meilleur".

2.10. Série corrigée des variations saisonnières (CVS)

La série corrigée des variations saisonnières est :

$Y(t) - S_a(t)$ dans le cas du modèle additif et

$Y(t)/S_m(t)$ dans le cas du modèle multiplicatif.

3. Modèle probabilisé associé à une expérience aléatoire et simulation

Programme de Première
(BO Hors Série n° 7 du 31 août 2000)

Modélisation d'expériences aléatoires de référence ; utilisation de modèles définis à partir de fréquences observées.

Simuler une expérience consiste à simuler un modèle de cette expérience.

↳ Premier exemple

- ***expérience aléatoire :***

on lance un dé cubique équilibré dont les six faces sont numérotées de 1 à 6

- ***ensemble des résultats possibles :***

$\Omega = \{1, 2, 3, 4, 5, 6\}$

- ***ensemble des événements :***

sous-ensembles de Ω

- ***probabilités :***

équiprobabilité : 1/6 pour chaque face

↳ Deuxième exemple

- ***expérience aléatoire :***

on lance un dé "pipé" dont les six faces sont numérotées de 1 à 6

- ***ensemble des résultats possibles :***

$\Omega = \{1, 2, 3, 4, 5, 6\}$

- ***probabilités :***

on lance 100 fois ce dé et on observe les fréquences de sortie de chaque face :
0.11 0.59 0.07 0.11 0.08 0.04

↳ Troisième exemple

- ***expérience aléatoire :***

on lance deux dés cubiques équilibrés dont les six faces sont numérotées de 1 à 6.

- ***ensemble des résultats possibles :***

Ω = {"1,1", "1,2", "1,3", "1,4", "1,5", "1,6", "2,2", "2,3", "2,4", "2,5", "2,6", "3,3", "3,4", "3,5", "3,6", "4,4", "4,5", "4,6", "5,5", "5,6", "6,6"}
soit 21 résultats,

ou bien

Ω' = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)} (c'est-à-dire $\{1,2,3,4,5,6\}^2$)
soit 36 résultats ?

- ***probabilités :***

équiprobabilité sur Ω ? équiprobabilité sur Ω' ?

Réponses : équiprobabilité sur Ω'
argument de symétrie
argument fréquentiste

↳ Quatrième exemple

À la fin du XVIème siècle, le Duc de Toscane aurait posé à Galilée le problème suivant au sujet du jeu de Passe-Dix.
Ce jeu consiste à lancer trois dés et à observer la somme des points marqués.

« Au jeu de Passe-Dix, on a 6 manières possibles de faire le 11 comme le 12. En effet :

Somme = 11 : "6,4,1" "6,3,2" "5,5,1" "5,4,2" "5,3,3" "4,4,3"

Somme = 12 : "6,5,1" "6,4,2" "6,3,3" "5,5,2" "5,4,3" "4,4,4"

Pourtant, lorsqu'on joue un grand nombre de fois, on s'aperçoit que le 11 sort plus souvent que le 12. Qu'en pensez-vous ? »

Solution : Equiprobabilité sur les 216 triplets de $\{1,2,3,4,5,6\}^3$

Evénement	**Probabilité**
trois chiffres égaux "4, 4, 4" $\rightarrow$ {(4, 4, 4)}	1/216
deux chiffres égaux et le 3ème distinct "5, 5, 2" $\rightarrow$ {(5, 5, 2), (5, 2, 5}, (2, 5, 5)}	3/216
trois chiffres distincts "5, 4, 3" $\rightarrow$ {(5, 4, 3), (5,3,4), (4,3,5), (4,5,3), (3,4,5), (3,5,4)}	6/216
somme = 11 "6,4,1" "6,3,2" "5,5,1" "5,4,2" "5,3,3" "4,4,3"	27/216
somme = 12 "6,5,1" "6,4,2" "6,3,3" "5,5,2" "5,4,3" "4,4,4"	25/216

Quatrième exemple, suite

Simulation du jeu de Passe-dix sur un tableur

- ***Définition de la simulation***

La simulation consiste à construire un échantillon de valeurs $(x_1,\ldots,x_K)$, qui puissent être considérées comme des observations de variables aléatoires réelles (v.a.r.) $(X_1,\ldots,X_K)$, indépendantes et identiquement distribuées (i.i.d.) comme une v.a.r. *X*, la distribution de *X* étant connue.
On parlera donc du nombre *K* de simulations.

- ***Objectif de la simulation***

L'objectif de la simulation est d'obtenir une approximation de la distribution de probabilité d'une variable *Y* fonction de *X* (ou de $X_1,\cdots,X_n$, *n* v.a.r. i.i.d. comme *X*) qu'il serait difficile d'obtenir par un calcul formel.
On remplace le calcul formel par du calcul numérique, la simulation consiste à produire les données adéquates.

Simulation du jeu de Passe-Dix sur un tableur

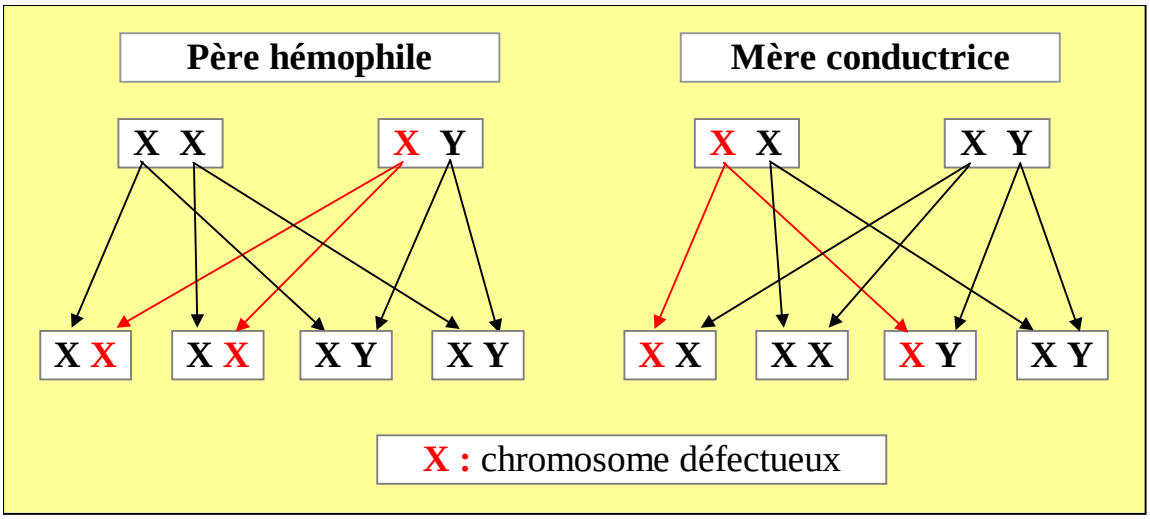
On utilise un générateur de nombres entiers aléatoires compris entre 1 et 6 (observation d'une v.a.r. *X* uniforme sur {1, 2, 3, 4, 5, 6}).

On utilise trois fois ce générateur (observation de trois v.a.r. X_1, X_2, X_3, i.i.d. comme *X*) et on calcule la somme des trois nombres générés
(observation d'une v.a.r. $Y = X_1 + X_2 + X_3$ à valeurs dans {3, 4, …, 18}).

On répète 10 000 fois cette expérience (d'où une série statistique de 10 000 observations de la v.a.r. *Y*) et on construit la distribution de fréquences de la variable *Y*, c'est-à-dire, la fréquence d'apparition sur les 10 000 expériences, des différentes valeurs {3, 4, …, 18}.

↳ Cinquième exemple. Modélisation de l'hémophilie.

Dans "Le Monde" du 20 avril, on trouve le rectificatif suivant concernant l'hémophilie, maladie génétique du sang.
« Le gène défectueux qui provoque cette affection est porté par le chromosome *X*. Lorsque le père est hémophile, toutes ses filles sont "conductrices", car elles héritent à la fois du chromosome *X* paternel (porteur de la tare) et d'un chromosome *X* maternel, mais aucun de ses fils ne sera atteint. Lorsque la mère est conductrice, un garçon sur deux ayant hérité du chromosome *X* maternel porteur de la tare est hémophile et une fille sur deux est conductrice. Un seul des deux chromosomes *X* des femmes est, en effet, atteint, l'autre étant normal. »



Comment modéliser ce phénomène ? Comment interpréter la dernière phrase ?
Une fille est conductrice. Quelle est la probabilité que sa mère soit conductrice ?
Dans la population y a-t-il autant d'hommes hémophiles que de femmes conductrices ?

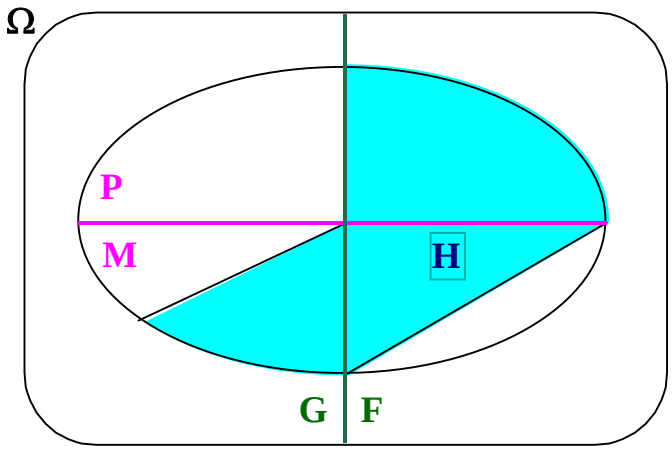
Convention
On nomme "hémophile" toute personne porteuse du chromosome défectueux.

Ω : population d'enfants

G : garçons
F : filles

P : enfants dont le père est hémophile
M : enfants dont la mère est hémophile
H : enfants hémophiles

$G \cap F = \varnothing$ et $G \cup F = \Omega$
$P \cap M = \varnothing$ maladie rare
$H \subset P \cup M$ maladie héréditaire
$G \cap H \subset M$ tous les garçons hémophiles ont une mère hémophile
$F \cap P \subset H$ toutes les filles d'un père hémophile sont hémophiles



On choisit un enfant "au hasard" dans cette population.
P est l'équiprobabilité sur Ω.

Traduction des hypothèses

$$\mathsf{P}(\mathrm{F}) = \mathsf{P}(\mathrm{G}) = \frac{1}{2}$$

$$\mathsf{P}_{\mathrm{G}\cap\mathrm{P}}(\mathrm{H}) = 0 \qquad \mathsf{P}_{\mathrm{G}\cap\mathrm{M}}(\mathrm{H}) = \frac{1}{2} \qquad \mathsf{P}_{\mathrm{F}\cap\mathrm{P}}(\mathrm{H}) = 1 \qquad \mathsf{P}_{\mathrm{F}\cap\mathrm{M}}(\mathrm{H}) = \frac{1}{2}$$

$$\mathsf{P}_{\mathrm{P}}(\mathrm{F}) = \mathsf{P}_{\mathrm{M}}(\mathrm{F}) = \mathsf{P}(\mathrm{F}) = \mathsf{P}(\mathrm{G}) = \mathsf{P}_{\mathrm{P}}(\mathrm{G}) = \mathsf{P}_{\mathrm{M}}(\mathrm{G}) = \frac{1}{2}$$

(indépendance entre le sexe de l'enfant et la transmission de la maladie)

$$\begin{aligned}
P(H \cap G) &= P(H \cap G \cap P) + P(H \cap G \cap M) \\
&= P_{G \cap P}(H)\, P(G \cap P) + P_{G \cap M}(H)\, P(G \cap M) \\
&= \frac{1}{2} P(G \cap M) = \frac{1}{2} P_M(G)\, P(M) = \frac{1}{4} P(M)
\end{aligned}$$

$$P(H \cap G) = \frac{1}{4} P(M)$$

$$\begin{aligned}
P(H \cap F) &= P(H \cap F \cap P) + P(H \cap F \cap M) \\
&= P_{F \cap P}(H)\, P(F \cap P) + P_{F \cap M}(H)\, P(F \cap M) \\
&= P(F \cap P) + \frac{1}{2} P(F \cap M) = P_P(F)\, P(P) + \frac{1}{2} P_M(F)\, P(M) \\
&= \frac{1}{2} P(P) + \frac{1}{4} P(M)
\end{aligned}$$

$$P(H \cap F) = \frac{1}{2} P(P) + \frac{1}{4} P(M)$$

Hypothèse

Les proportions d'hommes et de femmes parmi les hémophiles sont les mêmes d'une génération à la suivante.

$$\frac{P(H \cap F)}{P(H \cap G)} = \frac{P(M)}{P(P)} \quad \text{ou} \quad \frac{P(H \cap F)}{P(M)} = \frac{P(H \cap G)}{P(P)} \quad (=\alpha)$$

$$\begin{cases} P(H \cap G) = \alpha\, P(P) = \dfrac{1}{4} P(M) \\ P(H \cap F) = \alpha\, P(M) = \dfrac{1}{2} P(P) + \dfrac{1}{4} P(M) \end{cases} \Rightarrow \alpha = \frac{1}{2}$$

Conclusion :

$P(M) = 2\,P(P)$	il y a deux fois plus de femmes hémophiles que d'hommes hémophiles
$P_H(F) = 2\,P_H(G)$	parmi les enfants hémophiles, il y a deux fois plus de filles que de garçons $P_H(F) = \frac{2}{3}$ $\quad$ $P_H(G) = \frac{1}{3}$
$P_F(H) = 2\,P_G(H)$	la probabilité pour une fille d'être hémophile est deux fois celle pour un garçon d'être hémophile
$P_{H \cap F}(M) = \frac{1}{2}$	la probabilité pour une fille hémophile d'avoir une mère hémophile est $1/2$.

En conclusion …

1. La modélisation fait partie des mathématiques.
2. Une bonne connaissance de la statistique et du calcul des probabilités permet de modéliser les phénomènes les plus divers.
3. Il est possible, dans le cadre des programmes du lycée, de sensibiliser les élèves à la démarche de modélisation.
4. Les différentes étapes : expérimentation, observations, modélisation, simulation, …, nécessitent de la rigueur pour ne pas perdre de vue la structure mathématique qui se cache derrière la forêt des données numériques…