Introduction à la science des données

Micha Hersch

7 juillet 2025

version: a053c3c

Document produit sans intelligence artificielle

Chapitre 1

Les données

1.1 Introduction

1.1.1 Types de données

On peut distinguer différents types de données, selon leur caractéristiques.

— les *données catégorielles* représentent des catégories différentes, mais qui ne peuvent pas être ordonnées entre elles, par exemple la race d'un chien.

— Les *données ordinales* sont faites de valeurs qui peuvent être ordonnées entre elles, par exemple la satisfaction vis-à-vis d'un service (bonne, moyenne ou mauvaise). Ce sont des catégorie, mais on sait que "moyenne" est entre "bonne" et "mauvaise".

— Les données numériques sont représentées par des nombres. Parmi celles-ci, on distingue les *données numériques discrètes*, qui prennent leurs valeurs dans une ensemble discret, des *données numériques continues* qui prennent leurs valeurs dans un intervalle continu.

Les outils d'analyse de données seront différents selon le type de données considérées. Nous nous intéresserons ici principalement aux données numériques qui permettent d'utiliser de nombreux outils issus des mathématiques.

1.2 Données unidimensionnelles

Il est souvent utile de résumé un jeu de données en quelques valeurs clés permettant de se faire une idée du jeu de données sans avoir a lister toutes les valeurs.

1.2.1 Positionnement

Supposons que l'on souhaite résumer une série de données x_i ($i \in \{1, \ldots, n\}$) avec une seule valeur qui serait comme un résumé de ces données. Plusieurs pos-

sibilités peuvent être considérées.

La moyenne est calculée en divisant la somme des valeurs par le nombres de valeurs :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

La médiane est la valeur telle que la moitiés des données se trouve en dessous et l'autre moitié en dessus. Si le nombre de valeurs est pair, alors on prend le centre des deux valeurs du milieu. Pour trouver cette valeur il faut ordonner les x_i dans l'ordre croissant, pour obtenir des $x_{(i)}$ (avec une parenthèse !) et prendre la valeur suivante.

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impair} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{si } n \text{ est pair} \end{cases} \tag{1.2}$$

Le mode est la valeur qui est le plus souvent représentée.

Selon le type de données concernée, il sera possible ou pas de calculer de telles valeurs. Ainsi, la moyenne ne peut se calculer que sur des données numériques, alors que la médianes peut se calculer également sur des données ordinales, et le mode sur des données catégorielles.

Exemple 1.1 : Equipe de foot La taille (en cm) des joueuses d'une équipe de foot est donnée par le tableau (trié) suivant :

joueuse i	1	2	3	4	5	6	7	8	9	10	11
taille x_i	152	155	158	161	163	163	164	166	167	169	170

joueuse i	12	13	14	15
taille x_i	171	173	174	176

La *moyenne* est donnée par

$$\frac{1}{15}(154+155+158+161+163+163+164+166+167+169+170+171+173+174+176) = 165.6$$

La *médiane* est donnée par valeur du milieu, dans notre cas $x_8 = 166$ car on a 15 valeurs.

Quand au *mode* il s'agit de la valeur la plus fréquente, c'est-à-dire 163.

Le choix d'utiliser la moyenne, la médiane, ou le mode dépend du type des données (catégoriel, ordinal, numérique), mais aussi de l'utilisation qu'on souhaite faire de cette valeur et des caractéristiques des données. Par exemple, la moyenne est beaucoup plus sensible que la médiane aux valeurs extrêmes. En effet, les valeurs individuelles n'influencent la médiane que dans la mesure où elles sont supérieures ou inférieures à celle-ci.

Exemple 1.1 (suite) Dans l'équipe de foot, si deux très grandes nouvelles joueuses mesurant chacune 2 mètres arrivent dans l'équipe, la moyenne des tailles augmentera de 4 cm pour passer à 169.6 cm alors que la médiane n'augmentera que de 1 cm à 167 cm.

Exercice 1.1 Pour les situations suivantes, indiquer si la médiane ou la moyenne serait plus approprié comme représentant d'un jeu de donnée.

(a) les notes d'un élèves pour déterminer de sa promotion

(b) le salaire d'une population pour déterminer un seuil de droits à des allocations

(c) la consommation quotidienne d'alcool d'une personne

(d) le temps passé chaque jour par une personne sur les réseaux sociaux

(e) la quantité de CO2 émise par année pour tous les habitants de la planète

1.2.2 Dispersion

Bien qu'il soit utile d'avoir un nombre qui permette de savoir où se trouve approximativement les données, par exemple la moyenne ou la médiance, c'est aussi intéressant d'avoir un nombre qui indique la dispersion, c'est-à-dire à quel point les valeurs sont différentes les unes des autres. A nouveau, plusieurs mesures de dispersion existent :

La variance v est la moyenne des carrés des écarts à la moyenne. Autrement dit, on prend la différence de chaque valeur avec la moyenne, on l'élève au carré, et on prend la moyenne des valeurs ainsi obtenues. Mathématiquement, on obtient :

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1.3}$$

De la variance, on peut obtenir l'écart-type qui est simplement $\sqrt{\mathrm{Var}(x)}$. L'intérêt de l'écart-type est qu'il garde la même unité que les x_i (et $\bar{x}$). Si nos données correspondent à des cm, comme dans l'exemple 1.1, alors l'écart-type est aussi en cm, tandis que la variance correspond à des cm^2.

L'étendue est simplement la différence entre la valeur maximale est la valeur minimale, autrement dit $\max_i(x_i) - \min_i(x_i)$.

La distance interquartile est la différence entre le troisième et le première quartile des données. Si les données $x_{(i)}$ sont ordonnées dans l'ordre croissant, alors le premier quartile correspond à la valeur $x_{(\frac{n}{4})}$ telle que le quart des donnée y est inférieur. Le troisième quartile est la valeur $x_{(\frac{3n}{4})}$ telle que le quart des données y est supérieur. La distance interquartile $x_{(\frac{n}{4})} - x_{(\frac{n}{4})}$ est en fait l'étendue appliquée à la moitié des données se trouvant le plus au milieu. Cela permet à cette mesure d'être peu influencée par les valeurs extrêmes.

Exercice 1.2 Calculer la variance, l'étendue et la distance interquartile de l'exemple 1.1.

Exercice 1.3 On considère un jeu de données x_i. En comparant deux à deux l'écart-type, l'étendue et la distance interquartile, peut-on dire qu'une valeur est forcément plus grande que l'autre ?

Exercice 1.4 Démontrer l'égalité suivante

$$\mathrm{Var}(x) = \overline{x^2} - \bar{x}^2 \,, \tag{1.4}$$

où $\overline{x^2}$ est la moyenne des carré des x_i :

$$\overline{x^2} = \frac{1}{n} \sum_i^n x_i^2 \tag{1.5}$$

1.2.3 Distribution empirique

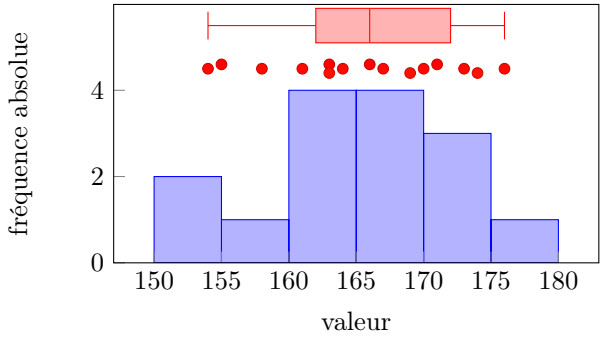
La distribution empirique d'un jeu de donnée consiste simplement à compter la fréquence des différentes valeurs obtenues. On peut considérer ces fréquences soit en valeurs absolues (c'est-à-dire le nombre d'observation correspondant à chaque valeur), soit en pourcentages des valeurs récoltées. Pour les représenter, on utilise souvent un histogramme, mais on peut également utiliser un tableau de valeurs.

Si les valeurs sont continues on va les classer dans des intervalles et compter le nombres de valeurs qui tombent dans chaque intervalle.
La distribution empirique est parfois résumée graphiquement sous forme de boîte à moustache (ou *boxplot* en anglais). Ce graphique représente un boite allant du premier au troisièmes quartiles, dans laquelle la position de la médiane est représentée par un trait, ainsi que des "moustaches" couvrant le reste de l'étendue de la distribution. Ces boites à moustaches sont représentées verticalement ou horizontalement. Diverses versions existent qui incluent parfois davantage d'informations telles que la moyenne, les valeurs extrêmes, ou des intervalles de confiance.

Exemple 1.1 (suite) Dans l'exemple de l'équipe de foot, on peut par exemple faire des intervalle de 5 cm et obtenir la distribution empirique suivante :

intervalles	$]150, 155]$	$]155, 160]$	$]160, 165]$	$]165, 170]$	$]170, 175]$	$]175, 180]$
fréquence	2	1	4	4	3	1

L'histogramme correspondant et représenté en bleu ci-dessous. La boite à moustaches (en rouge) résume cette distribution empirique en indiquant les quartiles. Les données ont également été représentées par des points pour faciliter la comparaison.



De manière générale, l'histogramme est ce qui est le plus informatif, mais c'est difficile à utiliser lorsqu'on veut comparer plusieurs distribution entre elles. Dans le cas, les boites à moustaches, que l'on peut facilement représenter côte-à-côte sont davantage utilisées.

1.3 Données multidimensionnelles

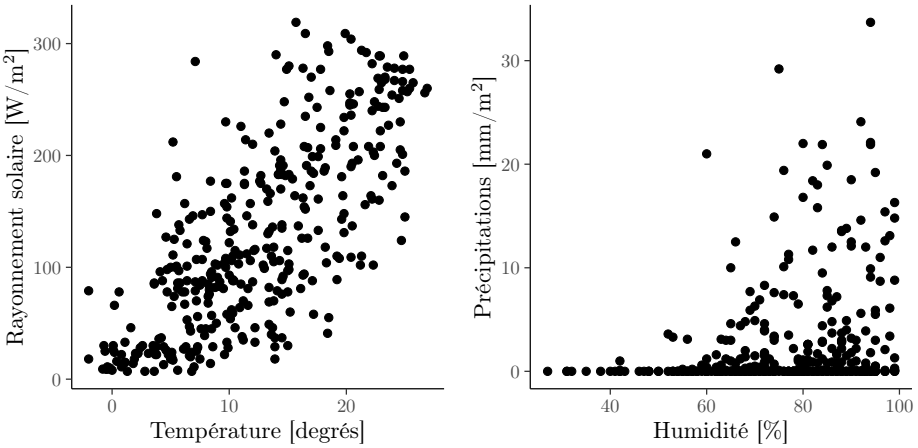
Un jeu de données contient d'habitude plusieurs informations sur le même objet. Ainsi les données ne peuvent pas être présentées comme une simple liste de valeurs dans dans l'exemple 1.1, mais sous forme de tableau à plusieurs colonnes, où chaque colonne représente une variable. On parle alors de jeu de données multidimensionnel.

Exemple 1.2 : La station météo d'Aigle On considère les enregistrements journaliers de la station météorologique d'Aigle. [1]Pour chaque jour, on extrait plusieurs mesures : température, humidité, précipitation, rayonnement solaire. On peut représenter ces données sous la forme d'un tableau où chaque ligne représente une observation, c'est-à-dire un jour dans notre cas, et chaque colonne une variable.

1. Les données sont disponibles sur `https://www.agrometeo.ch`

date	température [°C]	précipitation [mm/m^2]	humidité [%]	rayonnement solaire [W/m^2]
01.01.2024	4.6	0.1	84	26
02.01.2024	6.2	4.9	70	27
03.01.2024	9.9	4.8	68	28
04.01.2024	7.3	1.2	79	29
...	...	...	...	...
31.12.2024	-0.8	0	99	9

Ce jeu de donnée est de dimension 5 et contient 366 observations (2024 étant bissextile), correspondant au 366 lignes du tableau. A part la date (qui est une valeur ordinale facilement convertible en un entier), ce sont toutes des variable numériques continues, et donc on peut utiliser les mesures de dispersion telles que la variance. On ne peut pas représenter les cinq dimensions sur un même graphique, mais on peut les représenter deux à deux, par exemple comme ci-dessous



1.3.1 Covariance et corrélation

Le but d'analyser des données multidimensionnelles est souvent d'étudier les relation entre différentes variables. Une mesure souvent utilisée est la *covariance* entre deux variables. Pour ceci, on considère deux vecteurs $\vec{x}$ et $\vec{y}$ correspondant chacun à une colonne du tableau de données multidimensionnelles. Il ont la même longueur n correspondant aux nombres d'observations effectuées. De plus, on suppose pour l'instant que ces vecteurs sont centrés, c'est-à-dire que la moyenne de chacun d'entre eux est nulle : $\bar{x} = \bar{y} = 0$, où $\bar{x}$ est la moyenne des n éléments de $\vec{x}$ et $\bar{y}$ est la moyenne des n éléments de $\vec{y}$. La covariance est alors

donnée par la formule suivante :

$$\mathrm{Cov}(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{1.6}$$

Comme son nom l'indique, cette valeur mesure comment les valeurs de $\vec{x}$ et $\vec{x}$ co-varient. Elle peut être vue comme le produit scalaire de $\vec{x}$ et $\vec{y}$ (divisé par le nombre de dimension). Pour chaque observation x_i, y_i, on va regarder comment elle se situe par rapport à la moyenne des observations qui est nulle. Si les deux valeurs sont positives, donc supérieures à leur moyenne respective, alors le terme $x_i y_i$ est positif et va donc augmenter la covariance. De même si les deux valeurs sont négatives, $x_i y_i$ est positif. Par contre si x_i est positif et y_i est négatif, ou vice-versa, alors le terme $x_i y_i$ sera négatif et diminuera la valeur de la covariance. Autrement dit, la covariance sera positive si les x_i et les y_i ont en moyenne tendance est être ensemble supérieurs ou inférieurs à leur moyenne et elle sera négative s'ils ont tendance à être opposés l'un à l'autre (l'un est positif lorsque l'autre est négatif). Si aucune tendance ne se dessine parce que les variables prennent parfois des valeurs opposée et parfois pas, alors la covariance sera proche de 0.

Dans le cas où les valeurs ne sont pas centrées, c'est-à-dire que les moyennes des x_i et y_i ne sont pas nulles, alors il faut les centrer avant de calculer la covariance. On obtient alors la formule suivante, puisqu'on compare toujours les déviations par rapport à leur moyenne des x_i et y_i.

$$\mathrm{Cov}(\vec{x}, \vec{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{1.7}$$

La corrélation est une mesure permettant de quantifier par une valeur entre -1 et 1 à quel points deux variables sont co-déterminée. La mesure de corrélation la plus utilisée est la corrélation de Pearson ρ, qui correspond à la covariance normalisée par la produit des écarts-types :

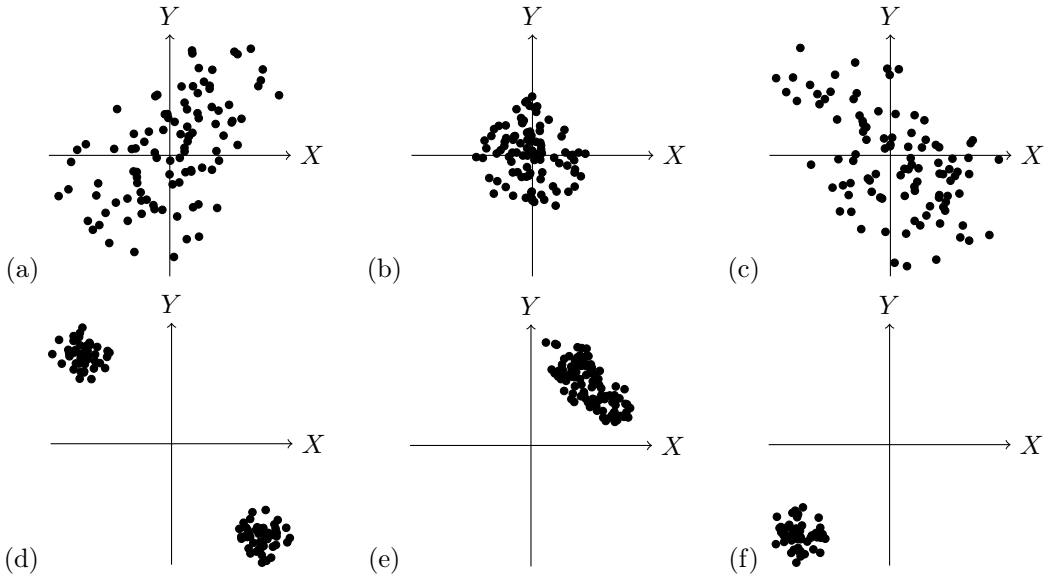
$$\rho(\vec{x}, \vec{y}) = \frac{\mathrm{Cov}(\vec{x}, \vec{y})}{\sqrt{\mathrm{Var}(\vec{x})\mathrm{Var}(\vec{y})}} \tag{1.8}$$

La corrélation de Pearson vaut 1 les points sont alignés sur une droite de pente strictement positive (et -1 si la pente est strictement négative). Un propriété importante de la corrélation que n'a pas la covariance est qu'elle ne dépend pas des unités utilisées.

Exemple 1.2 (suite) Dans l'exemple des données météorologiques d'Aigle, le rayonnement et la température ont une covariance positive qui vaut 438. Leur corrélation vaut 75%. La covariance et la corrélations entre l'humidité et les précipitation sont respectivement de 21 et 27%. Si on mesure les précipitations

en cm plutôt qu'en mm, la covariance est divisée par 10, mais la corrélation reste la même.

Exercice 1.5 On considère les données de deux variables représentées par les graphiques suivants (de même échelle). Ordonner les covariances correspondantes dans l'ordre croissant et en indiquer le signe.



Exercice 1.6 Calculer les corrélations suivantes :

(a) $\rho(\vec{x}, -\vec{x})$

(b) $\rho(\vec{x}, 2\vec{x})$

(c) $\rho(\vec{x}, 2\vec{x} + 3)$

(d) $\rho(\vec{x}, \vec{0})$

Exercice 1.7 Démontrer que

$$\mathrm{Cov}(x, y) = \overline{xy} - \bar{x}\bar{y}\,, \tag{1.9}$$

où

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} xy \tag{1.10}$$

1.3.2 Analyse en composantes principales

L'analyse en composante principale (ACP), en anglais *principal component analysis (PCA)*, permet de réduire la dimension d'un jeu de données en n'en conservant que quelques unes. Si l'on représente chaque observation (ou ligne du tableau de données) par un vecteur $\vec{x}_i$ dans un espace vectoriel à m dimensions, l'ACP consiste à projet ces vecteur dans un sous-espace de dimension inférieure à m. Pour choisir quel est ce sous-espace vectoriel, on va choisir celui pour lequel la variance des données est maximale. En effet, on peut faire l'hypothèse qu'une dimension dans laquelle les données ne varient presque pas n'est pas très intéressante à étudier. Ainsi, en supposant pour simplifier les calculs que les $\vec{x}_i$ sont centrés, il s'agit de trouver la direction qui maximise la variance des données, c'est-à-dire

$$\underset{\vec{v}}{\operatorname{argmax}} \sum_i (\vec{x}_i^T \vec{v})^2 \quad \text{s. c.} \quad \vec{v}^T \vec{v} = 1\,, \tag{1.11}$$

car $\vec{x}_i^T \vec{v}$ est la taille de la projection de $\vec{x}_i$ sur $\vec{v}$. En utilisant la méthode des multiplicateurs de Lagrange, il faut résoudre l'équation suivante en $\vec{v}$.

$$\frac{d}{d\vec{v}}\Big(\sum_i (\vec{x}_i^T \vec{v})^2\Big) - \lambda' \frac{d}{d\vec{v}}(\vec{v}^T \vec{v}) = \vec{0} \tag{1.12}$$

$$\Leftrightarrow \quad \frac{d}{d\vec{v}}\Big(\sum_i \vec{v}^T \vec{x}_i \vec{x}_i^T \vec{v}\Big) - 2\lambda' \vec{v} = \vec{0} \tag{1.13}$$

$$\Leftrightarrow \quad \frac{d}{d\vec{v}} \vec{v}^T \Big(\sum_i \vec{x}_i \vec{x}_i^T\Big) \vec{v} - 2\lambda' \vec{v} = \vec{0} \tag{1.14}$$

$$\Leftrightarrow \quad \frac{d}{d\vec{v}} (n-1) \vec{v}^T C \vec{v} - 2\lambda' \vec{v} = \vec{0} \tag{1.15}$$

$$\Leftrightarrow \quad C\vec{v} - \lambda \vec{v} = \vec{0}\,, \tag{1.16}$$

où C est la matrice de covariance des $\vec{x}_i$. Autrement dit, $\vec{v}$ est le vecteur propre de C, la matrice de covariance des $\vec{x}_i$. Ce résultat reste vrai si les $\vec{x}_i$ ne sont pas centrés, il faut simplement soustraire la moyenne des x_i à chaque apparition de x_i dans les équations ci-dessus qui restent ainsi valides.

Comme C est positive est symmétrique, on sait (par le théorême spectral) qu'elle contient m valeurs propres positives correspondant à m vecteurs propres orthogonaux. Les vecteurs propres correspondant à aux grandes valeurs propres correspondent aux directions dans lesquelles la variances est la plus grande. C'est donc sur ce sous-espace qu'il est intéressant de projeter les données. La somme des valeurs propres correspondant aux directions retenues indique le pourcentage de la variance généralisée qui est conservée par la projection.

Normalisation des données

Les composantes principales indiquent les directions dans laquelle la variance des données est maximisée. Toutefois cette dernière dépend fortement des uni-

tés choisies pour exprimer les valeurs. Par exemple, des données exprimées en mètres auront une variance bien supérieures à ces mêmes données exprimées en kilomètres. Ainsi, pour éviter les composantes principales ne soient déterminées par l'arbitraire des unités choisies pour chaque variable, il est souvent préférable de *normaliser* les données, c'est à dire en soustraire la moyenne, puis les diviser par leur écart-type. Autrement dit, chaque valeur x_i est remplacée par une valeur $\tilde{x}_i$ calculée ainsi :

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\mathrm{Var}(x)}}\,.$$

C'est ensuite sur les donnée $\tilde{x}_i$ et leur matrice de covariance que sont calculées les composantes principales.

1.4 Contrôle qualité

Le processus de récolte des données varient énormément selon les cas. Il peut être fait automatiquement (par exemples les traces numériques des internautes), être le résultats d'enquêtes (par exemple en marketing), impliquer des données personnelles qui doivent être anonymisées et pour lesquelles il faut obtenir le consentement éclairé des personnes (par exemple pour la recherche médicale), venir d'appareils très perfectionnés ou sensibles (pour des expériences de physique), etc. Dans tous les cas, il est important de s'assurer, autant que possible, que les données sont de qualité suffisante et qu'elle ne contiennent pas trop d'erreurs. Selon le contrôle que que l'on a sur le processus de récolte des données, on a divers possibilité. Par exemple, on peut répéter plusieurs fois une expériences scientifique pour vérifier que les données sont consistantes entre les deux expériences. Ou alors demander à plusieurs personnes de faire une enquête pour voir si elles obtiennent les mêmes résultats, ou faire plusieurs mesures, potentiellement à des moments différents et avec des appareils différents, et vérifier la cohérence des données.

Une fois la récolte terminée, on peut vérifier que les données correspondent bien à ce à quoi on s'attend. Notamment, on peut vérifier les éléments suivants :

1. La présence de valeurs manquantes
2. La présence de valeurs impossibles
3. La présence de données aberrantes (*outliers* en anglais), c'est-à-dire de données très différentes des autres

On peut également vérifier la présence de corrélations attendues, ou au contraire inattendues, entre les différentes variables ou entre les variable et certains paramètres de la récolte des données, par exemple la machine utilisée, la personne qui a pris le mesure, le moment de la mesure, etc.

1.5 Solutions des exercices

Exercice 1.3 L'étendue est plus grande que l'écart-type et la distance interquartile.

Exercice 1.5 $d < c < e < b \approx f \approx 0 < a$

Chapitre 2

Les modèles

2.1 Modélisation

En science des données, un modèle est une description mathématique d'un processus ou d'un phénomène réel ou fictif qui exprime une relation quantitative entre plusieurs valeurs. Un modèle exprime souvent cette relation comme une *valeur expliquée* déterminée en fonction d'une *valeur explicative*. Un modèle contient souvent des paramètres, généralement fixes, qui déterminent la relation.

Exemple 2.1 : Modèle de chute La position x d'un objet tombant depuis une certaine hauteur en fonction du temps t peut être donnée par le modèle suivant :

$$x = \frac{1}{2}gt^2 \tag{2.1}$$

où g est la constante gravitationnelle sur Terre.
Dans cet exemple, la position x est la variable expliquée, le temps t est la variable explicative, et la constante g est un paramètre du modèle.

Un modèle n'a pas forcément prétention à représenter exactement la réalité, il sera toujours une approximation visant à en capturer certains aspects. Dans l'exemple ci-dessus, le modèle ne tient pas compte du frottement de l'objet avec l'air, ni du fait qu'un fois arrivé à terre, il s'arrêtera. Ces aspects-là ne sont simplement pas représentés dans le modèle, ce qui peut être un inconvénient ou un avantage du modèle, selon l'utilisation qu'on souhaite en faire.

Exemple 2.1 (suite) Un autre modèle pour représenter le même phénomène pourrait être le suivant :

$$x = \begin{cases} vt & \text{si } t < \frac{h}{v} \\ h & \text{si } t \geq \frac{h}{v} \end{cases} \tag{2.2}$$

où v est la vitesse de chute et h la hauteur de chute. Ce modèle fait l'hypothèse que l'objet chute à vitesse constante (ce qui peut être raisonnable par exemple s'il a un parachute) et s'arrête lorsqu'il touche le sol.

Ainsi, on ne peut pas vraiment dire qu'un modèle est "juste" ou "faux", il sera simplement plus ou moins approprié de l'utiliser selon les situations. Une grande partie des débats scientifiques porte sur la pertinence, ou pas, des modèles utilisés.

2.2 Variable aléatoire

2.2.1 Définition

La définition et le traitement mathématiquement rigoureux des variables aléatoires dépassant le cadre de ce cours, nous considérons simplement qu'une variable aléatoire est une variable X qui peut prendre différentes valeurs possibles. L'ensemble Ω des valeurs possibles est appelé *l'univers* de X.
A chaque sous-ensemble E de Ω est associée une probabilité $P(X \in E)$ que la variable X lui appartient. Dans le reste de ce document, P(vnement) représente la probabilité de l'événement indiqué entre parenthèses.

Exemple 2.2 : Jet d'un dé Le résultat d'un jet d'un dé à six faces peut être décrit par une variable aléatoire X. L'univers des possibilités est donné par

$$\Omega = \{1, 2, 3, 4, 5, 6\} \tag{2.3}$$

La probabilité associée à chacune des possibilités est donnée par $\frac{1}{6}$.

$$P(X = i) = P(X \in \{i\}) = \frac{1}{6} \quad \forall i \in \Omega. \tag{2.4}$$

La probabilité d'avoir une nombre inférieur à 3 est donnée par

$$P(X < 3) = P(X \in \{1, 2\}) = \frac{2}{6} = \frac{1}{3}. \tag{2.5}$$

L'univers Ω d'une variable aléatoire peut être composé de nombres, mais ce n'est pas forcément le cas. Par exemple pour la variable aléatoire décrivant les

auteurs ou autrices de livres dans une bibliothèque, l'univers est constitué de tous les noms possibles.
Si l'ensemble Ω est constitué de valeurs numériques discrètes, comme dans l'exemple ci-dessus, la variable aléatoire est dite *discrète*. A l'inverse, si Ω est constitué de valeurs numériques continues, par exemple le poids d'une pomme, la variable aléatoire est dite *continue*.

Exercice 2.1 Donner l'univers et, si vous le pouvez, les probabilités associées d'une variable aléatoire décrivant les phénomènes suivants :

(a) la mesure de la taille d'une personne

(b) une personne choisie au hasard dans la classe

(c) la couleur d'une carte tirée au hasard dans un jeu standard de 36 cartes.

2.2.2 Distribution

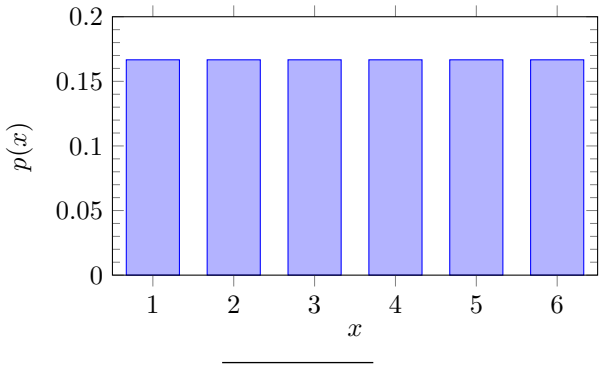
La *distribution de probabilité* (ou *loi de probabilité*) d'une variable aléatoire décrit la probabilité associée à chaque valeur de son univers Ω. Dans le cas d'une variable aléatoire numérique, cette distribution peut prendre la forme une fonction $p : \Omega \to \mathbb{R}_+$. Pour qu'il s'agisse bien d'une distribution de probabilité, cette fonction p doit satisfaire la condition suivante car il est certain que X fait partie de l'univers Ω :

$$\sum_{x \in \Omega} p(x) = 1 \quad \text{si } X \text{ est discrète,} \tag{2.6}$$

$$\int_{x \in \Omega} p(x)dx = 1 \quad \text{si } X \text{ est continue.} \tag{2.7}$$

Dans le cas où cette variable aléatoire est discrète, on peut représenter cette fonction sous forme d'histogramme dont la somme des valeurs vaut 1. Si elle est continue, on peut la représenter sous forme de courbe dont l'aire sous la courbe vaut 1.

Exemple 2.2 (suite) Dans l'exemple du jet d'un dé, la loi de probabilité peut être représentée par l'histogramme suivant :



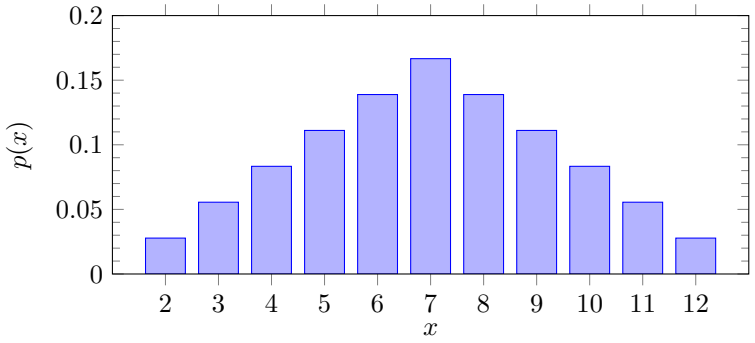
Exemple 2.3 : Jet de deux dés On jette deux dés à six face et on calcule la somme des nombres obtenus. On représenter les possibilités de la manière suivante ou le résulat du premier dé est donné par la colonne et celui du second dé est donné par la ligne.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

La loi de probabilité correspondante peut être donnée par le tableau suivant où les probabilités sont obtenues en calculant la proportion de chacune des valeurs représentées dans le tableau ci-dessus.

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Cette loi de probabilité peut aussi être représentée sous forme d'histogramme :



Dans le cas continu, p est appelé une *fonction de densité de probabilité*. On parle de densité car, contrairement au cas discret, $p(x)$ donne la probabilité relative d'une valeur x par rapport aux autres valeurs possibles, et pas la probabilité absolue qui est nulle. (Il y a une infinité de possibilités pour X donc la chance de tomber exactement sur x vaut 0.) Par contre, cette fonction est utile pour calculer la probabilité (absolue) que X soit dans un intervalle donnée $[a\ b]$: il s'agit de l'aire sous la courbe de p entre a et b.

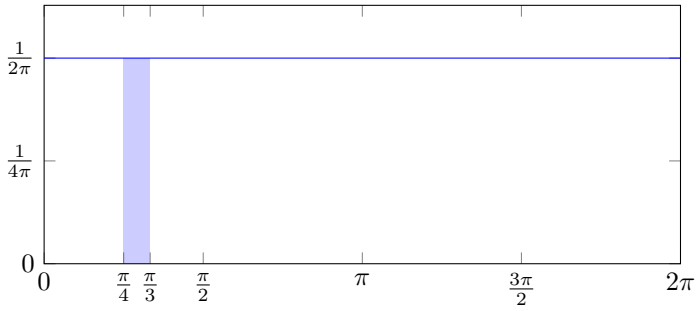
Exemple 2.4 : Allumette On lance au hasard une allumette parterre et on décrit par la variable aléatoire X l'angle que l'allumette fait avec le nord. L'univers Ω est donné par tous les angles possibles, c'est à dire l'intervalle compris entre 0 et 2π radians $\Omega = [0\ 2\pi]$. Comme l'allumette a une probabilité égale de tomber orientée dans n'importe quelle direction la fonction de densité de probabilité X vaut

$$p(x) = \frac{1}{2\pi}\,. \tag{2.8}$$

La probabilité que l'allumette tombe avec une orientation par rapport au nord comprise entre 45 et 60 degrés, c'est-à-dire entre $\frac{\pi}{4}$ et $\frac{\pi}{3}$ radian vaut

$$P(\frac{\pi}{4} \leq X < \frac{\pi}{3}) = (\frac{\pi}{3} - \frac{\pi}{4}) \cdot \frac{1}{2\pi} = \frac{4\pi - 3\pi}{12 \cdot 2\pi} = \frac{1}{24}$$

La situation peut se représenter graphiquement de la manière suivante, où cette probabilité correspond à l'aire en bleu.



Exercice 2.2 Reprendre l'exemple 2.4 et définir la variable aléatoire X comme l'angle entre l'allumette et le nord, mais cette fois en degrés. Exprimer alors la fonction de densité de probabilité de X et et l'utiliser pour calculer la probabilité que X prenne une valeur entre 45 et 60.

Exercice 2.3 En considérant l'exemple 2.4, quelle est la probabilité que l'allumette tombe avec un angle inférieur à $\frac{\pi}{10}$ par rapport au nord ?

Exercice 2.4 Soit une variable aléatoire continue X prenant des valeurs entre a et b. Soit $F(x)$ la fonction décrivant la probabilité $P(X < x)$.

(a) Que valent $F(a)$ et $F(b)$?

(b) Montrer de la fonction de densité de probabilité de $p(x)$ de X est la dérivée de $F(x)$: $p(x) = F'(x)$

La fonction $F(x)$ est appelée la *fonction de répartition* ou *fonction de distribution cumulative* de la variable aléatoire X et constitue une autre manière de la caractériser.

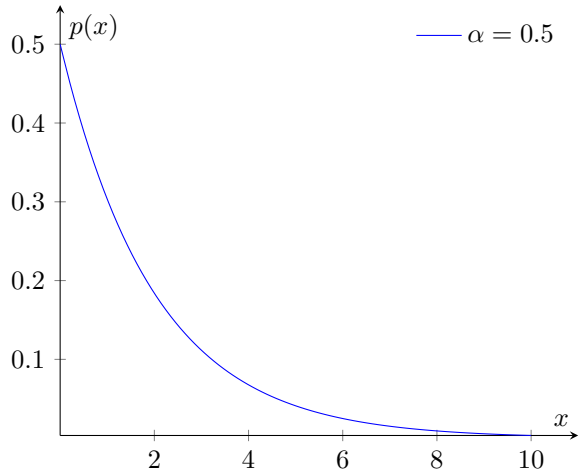
Exemple 2.5 : Le funambule Un funambule est en équilibre sur une corde. On suppose qu'à chaque petit intervalle Δt, il a une probabilité $\alpha\Delta t$ de tomber (et donc une probabilité $(1 - \alpha\Delta t)$ de ne pas tomber). On considère la variable aléatoire X décrivant le temps que le funambule reste sur la corde avant de tomber. La probabilité qu'il ne soit pas tombé après un temps x constitué de n intervalles Δt ($x = n\Delta t$) est donc $(1 - \alpha\Delta t)^n$. En passant à la limite $\Delta t \to 0$ (et donc $n \to \infty$), on obtient :

$$\begin{aligned}\lim_{\Delta t\to 0}(1-\alpha\Delta t)^n &= \lim_{n\to\infty}(1-\alpha\frac{x}{n})^n = \lim_{n\to\infty} e^{\log\left((1-\alpha\frac{x}{n})^n\right)} \\ &= \lim_{n\to\infty} e^{n\log(1-\alpha\frac{x}{n})} = \lim_{n\to\infty} e^{n(-\alpha\frac{x}{n})} = e^{-\alpha x}\,.\end{aligned}$$

La probabilité que le funambule soit tombé après un temps x est donc $1 - e^{-\alpha x}$. La densité de probabilité que le funambule tombe à l'instant x est donc la dérivée, c'est-à-dire

$$p(x) = \alpha e^{-\alpha x}\,. \tag{2.9}$$

Cette densité de probabilité, appelée la *distribution exponentielle*, peut être représentée par le graphe suivant (pour $\alpha = 0.5$).



On remarque que le funambule a moins de chances de rester un long moment sur le fil, ce qui est logique car plus le temps avance, plus les chances de tomber s'accumulent.

Exercice 2.5 En reprenant l'exemple 2.5, calculer les probabilités suivantes :

(a) La probabilité que le funambule reste moins de 5 secondes sur la corde s'il a une chance sur mille de tomber la première miliseconde.

(b) La probabilité que le funambule reste entre 30 et 40 secondes sur la corde s'il a une chance sur 100 de tomber la première seconde.

(c) La probabilité que le funambule reste plus d'une minute sur la corde s'il a une chance sur 50 de tomber la première seconde.

Exercice 2.6 Vérifier que la fonction $p(x)$ de l'exemple 2.5 est bien une densité de probabilité. Est-ce la cas pour tout $\alpha \in \mathbb{R}$?

Exercice 2.7 Soit p une fonction décrivant la loi de probabilité d'une variable aléatoire X d'univers Ω. Est-ce que $p(x)$ peut être supérieur à 1 ? Traiter séparément les cas où X discrète et où X est continue.

2.2.3 Opérations

Diverses opérations peuvent être effectuées sur des variables aléatoires.

Espérance

L'espérance d'une variable aléatoire numérique $X \in \Omega$ indique (lorsqu'elle est définie) la valeur moyenne qui peut être attendue pour celle-ci. Si X prend des valeurs discrètes, son espérance $E(X)$ est donnée par :

$$E(X) = \sum_{x \in \Omega} p(x) \cdot x, \tag{2.10}$$

c'es-à-dire qu'on mutliplie chaque valeur possible par sa probabilité et on somme le tout. Si X prend des valeurs continues, on ne peut pas les énumérer, donc on utilise le pendant différentiel de l'équation ci-dessus :

$$E(X) = \int_{x \in \Omega} p(x) \cdot x dx, \tag{2.11}$$

L'espérance d'une variable aléatoire est analogue au centre de gravité de son univers Ω si chaque point de l'univers avait une masse proportionnelle à sa probabilité.

Exemple 2.3 (suite) L'espérance de l'exemple du jet des deux dés, est donnée par

$$E(X) = \sum_{x\in\Omega} p(x)\cdot x = 2\cdot\frac{1}{36} + 3\cdot\frac{1}{18} + 4\cdot\frac{1}{12} + 5\cdot\frac{1}{9} + 6\cdot\frac{5}{36} + 7\cdot\frac{1}{6} + 8\cdot\frac{5}{36} + 9\cdot\frac{1}{9} + 10\cdot\frac{1}{12} + 11\cdot\frac{1}{18} + 12\cdot\frac{1}{36} = 7$$

Exemple 2.4 (suite) L'espérance de l'exemple de l'allumette est donnée par

$$E(X) = \int_{x\in\Omega} p(x)\cdot x dx = \int_0^{2\pi} \frac{1}{2\pi}\cdot x = \left[\frac{x^2}{4\pi}\right]_0^{2\pi} = \frac{4\pi^2}{4\pi} - 0 = \pi$$

Exercice 2.8 Calculer l'espèrance de X dans l'exemple 2.4, si X avait été défini entre $-\pi$ et π au lieu d'entre 0 et 2π. Est-ce que l'espérance nous donne une indication sur l'orientation "moyenne" de l'allumette ?

Exercice 2.9 En reprenant l'exemple 2.5, donner l'intégrale qui exprime l'espérance de X, c'est à dire le temps moyen que le funambule reste sur la corde avant de tomber. Tenter de résoudre cette intégrale en utilisant la technique de l'intégration par parties.

Variance

La variance $\mathrm{Var}(X)$ d'une variable aléatoire numérique $X \in \Omega$ indique à quel point cette variable s'éparpille autour de son espérance. Elle est donnée par l'espérance des carrés des écarts à la l'espérance de X (ou la moyenne des carrés des écart à la moyenne de X) :

$$\mathrm{Var}(X) = \sum_{x\in\Omega} p(x)\cdot\big(x - E(X)\big)^2, \quad \text{si } X \text{ est discrète,} \tag{2.12}$$

$$\mathrm{Var}(X) = \int_{x\in\Omega} p(x)\cdot\big(x - E(X)\big)^2 dx, \quad \text{si } X \text{ est continue.} \tag{2.13}$$

Exemple 2.3 (suite) La variance de X dans l'exemple du jet des deux dés, est donnée par

$$\begin{aligned}\mathrm{Var}(X) = \sum_{x\in\Omega} p(x)\cdot\big((x-E(X)\big)^2 &= (2-7)^2\cdot\frac{1}{36}+(3-7)^2\cdot\frac{1}{18}+(4-7)^2\cdot\frac{1}{12}\\ &+(5-7)^2\cdot\frac{1}{9}+(6-7)^2\cdot\frac{5}{36}+(7-7)^2\cdot\frac{1}{6}+(8-7)^2\cdot\frac{5}{36}\\ &+(9-7)^2\cdot\frac{1}{9}+(10-7)^2\cdot\frac{1}{12}+(11-7)^2\cdot\frac{1}{18}+(12-7)^2\cdot\frac{1}{36}=\frac{35}{6}\end{aligned}$$

Exemple 2.4 (suite) Dans l'exemple de l'allumette, la variance est donnée par

$$\begin{aligned}E(X) = \int_{x\in\Omega} p(x)\cdot\big(x-E(X)\big)^2dx &= \int_0^{2\pi}\frac{1}{2\pi}\cdot(x-\pi)^2dx = \frac{1}{2\pi}\int_0^{2\pi}x^2-2\pi x+\pi^2dx\\ &= \frac{1}{2\pi}\left[\frac{x^3}{3}-\pi x^2+\pi^2x\right]_0^{2\pi} = \frac{1}{2\pi}\big(\frac{8\pi^3}{3}-4\pi^3+2\pi^3\big)=\frac{\pi^2}{3}\end{aligned}$$

Exercice 2.10 On définit $E(X^2)$ comme l'espérance du carré de X, c'est-à-dire la moyenne des X^2.

(a) Comparez $E(X^2)$ et $E(X)^2$, c'est à dire le carré de l'espérance. Peut-on dire que l'un est plus grand que l'autre, et le cas échéant lequel ? Justifier sa réponse.

(b) En utilisant la définition de l'expérance, exprimer mathématiquemment $E(X^2)$ pour les cas discret et continu.

(c) En développant la définition de la variance ci-dessus, démontrer pour le cas discret que $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Opérations scalaires

Une variable aléatoire X peut être multipliée par un scalaire k. Cela signifie que toutes les valeurs prises par X sont mutlitpliée par le nombre k qui est constant. Le résultat Y est une nouvelle variable aléatoire dépendante de X :

$$Y = k\cdot X. \tag{2.14}$$

De la même manière, une constante k peut être ajouté à une variable aléatoire X pour donner une variable aléatoire Y dépendante de X.

$$Y = X + k. \tag{2.15}$$

Exercice 2.11 En partant de leur définition, calculer, en fonction de $E(X)$ et $\mathrm{Var}(X)$, les valeurs suivantes :

(a) $E(X + k)$

(b) $\mathrm{Var}(X + k)$

(c) $E(kX)$

(d) $\mathrm{Var}(kX)$.

(e) En déduire $E(aX + b)$ et $\mathrm{Var}(aX + b)$.

Sommes et multiplication de variables aléatoires

Il est également possible d'additionner, multiplier, soustraire ou diviser des variables aléatoires entre elles. Il en résulte à chaque fois une autre variable aléatoire.

Exemple 2.3 (suite) La variable aléatoire X décrivant le résultat du jet de deux dés peut être vue comme la somme de deux variables aléatoires Y et Z correspondant chacune au jet d'un dé.

$$X = Y + Z \tag{2.16}$$

Le même principe peut être appliqué avec le produit, la soustraction ou la division des valeurs obtenues.

Exercice 2.12 On considère X et Y, deux variables aléatoires continues de distribution uniforme, c'est-à-dire dont la fonction de densité de probabilité est constante, entre 0 et 2.

(a) Exprimer $p(x)$, la fonction de densité de probabilité de X et Y.

(b) On considère la variable aléatoire $Z = X + Y$. Calculer sa fonction de densité de probabilité.
Indice: Il faut considérer deux cas, selon si $Z < 2$ ou pas, et intégrer sur la valeur prise par X (ou Y).

2.2.4 Distribution conditionnelle

Considérons deux variables aléatoires X, Y. La *distribution conditionnelle de X en fonction de Y* décrit la distribution de X connaissant la valeur prise par Y. Il s'agit d'une fonction dépendante de x et y qui se note $p(x|y)$. Si on est intéressé uniquement à une seule valeur y prise par Y, alors on obtient une fonction $p(x|Y = y)$, dépendant uniquement de x.

Exemple 2.6 Le tableau suivant décrit le nombres de personnes ayant survécu au naufrage du Titanic en fonction de la classe de leur billet.

	1^e classe	2^e classe	3^e classe	Total
Femmes	140/144	80/93	76/165	296/402
Hommes	57/175	14/154	75/387	146/716
Enfants	5/6	24/24	27/79	56/109
Total	202/325	118/271	178/631	498/1227

On peut décrire la survie d'un passager par une variable aléatoire X prenant les valeur S (survivant) ou V (victime), et sa classe par une variable aléatoire $Y \in \{1, 2, 3\}$. Sans connaître Y, la distribution de X peut être décrites par.

$$P(X = S) = 489/1227 \quad P(X = V) = 738/1227 \tag{2.17}$$

Si on connaît dans quelle classe la personne a voyagé, on peut avoir une idée plus précise de la survie d'une personne grâche à sa distribution conditionnelle.

		X	
	$p(X\|Y)$	S	V
Y	1	202/325	123/325
	2	118/171	53/171
	3	178/631	453/631

On remarque dans le tableau ci-dessus que la somme des lignes vaut 1. Les trois lignes du tableaux donnent respective $p(X|Y = 1)$, $p(X|Y = 2)$ et $p(X|Y = 3)$.

Exercice 2.13 On lance deux dés à six faces et la somme obtenue est représentée par la variable aléatoire X. En vous aidant du tableau de l'exemple 2.3, donner les lois de probabilités suivantes :

(a) $p(X|Y)$ où Y est le résultat obtenu par le premier dé.
(b) $p(X|Y = 3)$ où $Y = 3$ signifie que le premier dé vaut 3.
(c) $p(X|Y = 3)$ où $Y = 3$ signifie qu'un des deux dés vaut 3.
(d) $p(X|Y \geq 4)$ où $Y \geq 4$ signifie que le plus petit des deux dés est supérieur ou égal à 4.
(e) $p(Y|X = 8)$ où Y est le résultat obtenu par le premier dé.

Exercice 2.14 Reprendre l'exemple 2.6 et déterminer les distributions suivantes où X est la survie ou non de la personne, Y est la classe dans laquelle elle a voyagé et Z est la catégorie de personne (homme, femme ou enfant).

(a) $pX|Z)$
(b) $p(Z|X)$
(c) $p(Z|Y)$

2.2.5 Distribution conjointe

Une autre manière de combiner deux variables aléatoire X et Y consiste à considérer leurs distribution conjointe $p(X, Y)$, qui indique la probabilité que $X = x$ et $Y = y$.

Exemple 2.6 (suite) En reprenant l'exemple du *Titanic*, la probabilité conjointe entre la classe et la survie des passager est donnée par le tableau suivant :

		X	
	$p(X,Y)$	S	V
	1	202/1227	123/1227
Y	2	118/1227	53/1227
	3	178/1227	453/1227

On remarque dans le tableau ci-dessus que la sommes des valeurs du tableau vaut 1. On peut calculer la probabilité en prenant un passager ou une passagère au hasard, en comptant que le nombre de passager dans chaque classe.

Y	1	2	3
$p(Y)$	325/1227	171/1227	631/1227

On remarque alors que

$$
\begin{aligned}
P(X=S,Y=1) &= \frac{202}{1227} = \frac{202}{325}\cdot\frac{325}{1227} = P(X=S|Y=1)\cdot P(Y=1)\\
P(X=S,Y=2) &= \frac{118}{1227} = \frac{118}{171}\cdot\frac{171}{1227} = P(X=S|Y=2)\cdot P(Y=2)\\
P(X=S,Y=3) &= \frac{178}{1227} = \frac{118}{631}\cdot\frac{631}{1227} = P(X=S|Y=3)\cdot P(Y=3)\\
P(X=V,Y=1) &= \frac{123}{1227} = \frac{123}{325}\cdot\frac{325}{1227} = P(X=V|Y=1)\cdot P(Y=1)\\
P(X=V,Y=2) &= \frac{53}{1227} = \frac{53}{171}\cdot\frac{171}{1227} = P(X=V|Y=2)\cdot P(Y=2)\\
P(X=V,Y=3) &= \frac{453}{1227} = \frac{453}{631}\cdot\frac{631}{1227} = P(X=V|Y=3)\cdot P(Y=3)
\end{aligned}
$$

Cet exemple peut être généralisé à la *formule des probabilités composées* qui permet d'exprimer la relation entre la probabilité conditionnelle et la probabilité conjointe.

$$P(X,Y) = P(X|Y)\cdot P(Y) = P(Y|X)\cdot P(X) \tag{2.18}$$

Autrement dit, la probabilité de deux événements peut se calculer comme le produit de la probabilité de l'un d'entre eux multipliée par la probabilité conditionnelle de l'autre événement par rapport au premier.

La distribution

$$p(x) = \sum_{y \in \Omega_Y} p(x|y) \cdot p(y) \quad \text{si } Y \text{ est discret} \tag{2.19}$$

$$p(x) = \int_{y \in \Omega_Y} p(x|y) \cdot p(y) dy \quad \text{si } Y \text{ est continu} \tag{2.20}$$

est appelée la *distribution marginale* de X par rapport à Y. Elle décrit la probabilité de X si on ne connaît pas la valeur prise par Y mais seulement sa distribution.

Exercice 2.15 Reprendre l'exemple 2.6 et déterminer les distributions suivantes où X est la survie ou non de la personne, Y est la classe dans laquelle elle a voyagé et Z est la catégorie de personne (homme, femme ou enfant).

(a) $p(X, Z)$

(b) $p(X, Y, Z)$

(c) $p(X, Y|Z)$

(d) $p(X|Y, Z)$

Exercice 2.16 - Loi de Bayes En utilisant la formule des probabilités composées, démontrer la loi de Bayes :

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{2.21}$$

2.2.6 Indépendance

La notion d'indépendance entre deux variables aléatoires est fondamentale pour l'interprétation de données. Deux variables aléatoires X et Y sont dites *indépendantes* si les valeurs prises ne dépendent pas l'une de l'autre. En d'autre terme, la valeur prise par une des variables aléatoire ne donne aucune indication et n'a aucune influence sur la valeur prise par l'autre variable aléatoire. L'indépendance entre deux variables aléatories X et Y aux univers Ω_x et Ω_x peut s'exprimer mathématiquement par le symbole $\perp\!\!\!\perp$, que l'on peut définir ainsi

$$X \perp\!\!\!\perp Y \Leftrightarrow p(x|y) = p(x) \quad \text{et} \quad p(y|x) = p(y) \quad \forall x \in \Omega_x,\ \forall y \in \Omega_y\,. \tag{2.22}$$

Exemple 2.7 On lance deux dés, et les résultats obtenus pour chaque dé est représenté par une variable aléatoire. Dans ce cas, ces deux variables aléatoires sont indépendantes car le résultat affiché par un dé ne donne aucune indication sur le résultat affiché par l'autre dé.
Par contre, si on lance deux dés et on représente la plus grande valeur obtenue par une variable aléatoire et la plus petite par une autre variable, alors ces deux

variables aléatoires ne sont pas indépendantes. En effet, si on connait la valeur prise par la première, on sait que la valeur prise par la seconde sera inférieure (ou égale).

Exercice 2.17 Indiquer si les variables aléatoires X et Y suivantes sont indépendantes ou pas.

(a) On considère une cohorte de patients en Suisse romande, dont on représente la taille par X et le sexe par Y.

(b) Un institut de sondage interroge la population Suisse, et décrit commune de domicile des personnes sondées par X et leur positionnement politique par Y

(c) On lance deux dés à 6 faces. X représente la valeur du premier dé. Si $X \leq 3$, alors Y vaut la valeur du second dé. Sinon Y vaut 7 moins la valeur du second dé.

(d) On lance deux dés à 6 faces. X représente la valeur du premier dé. Si $X \leq 3$, alors Y vaut la valeur du second dé. Sinon Y vaut 6 moins la valeur du second dé.

En appliquant la formule des probabilités composées, on obtient que

$$X \perp\!\!\!\perp Y \Leftrightarrow p(x,y) = p(x|y) \cdot p(y) = p(x) \cdot p(y) \quad \forall x \in \Omega_x,\ \forall y \in \Omega_y\,. \tag{2.23}$$

En d'autres termes, pour toutes les valeurs de possibles pour X et Y, la probabilité d'obtenir les deux valeurs x et y vaut le produit des probabilités d'obtenir x et d'obtenir y.

Exemple 2.7 (suite) Si le résultat de chacun des dés est une variable aléatoire indépendante, la probabilité d'obtenir la value x pour le premier dé et la valeur y pour le second dé vaut.

$$p(x,y) = p(x) \cdot p(y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \quad \forall x \in \Omega_x,\ \forall y \in \Omega_y\,. \tag{2.24}$$

Par contre, si Y correspond au plus petit des deux nombres tirés, alors ce n'est pas le cas. Par exemple pour $X = 6$ et $Y = 5$.

$$\begin{aligned} P(X=6, Y=5) &= P(X=6|Y=5) \cdot P(Y=5) = \frac{2}{3} \cdot \frac{3}{36} = \frac{1}{18} \\ &\neq P(X=6) \cdot P(Y=5) = \frac{1}{6} \cdot \frac{3}{36} = \frac{1}{72} \end{aligned}$$

Exercice 2.18 Calculer les distributions de probabilité des variables aléatoires suivantes correspondante aux deux cas de l'exemple 2.7.

(a) $P(X, Y)$

(b) $P(X|Y)$

(c) $P(Y|X)$

(d) $P(X)$

(e) $P(Y)$

Indice : Vous pouvez vous aidez d'une matrice représentant toutes les valeurs possibles pour X et Y et leur probabilité.

Exercice 2.19 En partant de leur définition, calculer $E(X+Y)$ et $\mathrm{Var}(X+Y)$ si X et Y sont indépendantes.

Variables aléatoires i.i.d.

On parle souvent de variables aléatoire *indépendantes et identiquement distribuées* (ou i.i.d.) pour indiquer qu'elles sont tirées de la même distribution de probabilité et de manière indépendante l'une de l'autre. En présence d'un phénomène qui se répète, on peut soit considérer qu'il s'agit de plusieurs réalisation indépendantes de la même variable aléatoire X, soit qu'ils s'agit de plusieurs variables aléatoires différentes X_i indépendante et identiquement distribuées.

Exemple 2.8 2.7 On mesure le poids des pièces produites dans une usine. Même si les pièces sont théoriquement identiques, elles n'auront pas forcément toutes exactement le même poids, à cause de petites variations dans la productions. Par contre on pourra considérer toutes les mesures comme réalisations de variables aléatoires indépendantes et identiquement distribuées (ou comme réalisations indépendantes de la même variable aléatoire), car issues du même processus de production et de mesure.

Il s'agit souvent d'une hypothèse qui est faite pour faciliter l'analyse de données et qui est plus ou moins justifiée selon les cas.

Exemple 2.9 On mesure chaque jour la concentration de CO_2 dans une salle de classe. On peut faire l'hypothèse que ces mesures sont des valeurs prises par des variables aléatoires indépendantes et identiquement distribuées. Cette hypothèse n'est sans doute pas absolument correcte, vu que la concentration de CO_2 a tendance à augmenter avec le temps, et à être influencée par d'autres facteurs tels que le moment de la journée, la présence de personnes ou de plantes dans la pièce. Les valeurs obtenues seront donc sans doutes différentes entre la semaine et le week-end ou pendant les vacances. Selon le contexte, cela pourra

toutefois quand même être utile de faire cette hypothèse, mais il faudra garder en tête cette approximation de la réalité en interprétant les résultats.

Exercice 2.20 Dans les exemples suivants, indiquer les éléments qui peuvent justifier ou infirmer l'hypothèse de variables i.i.d.

(a) ...

(b) ...

(c) ...

(d) ...

[à compléter]

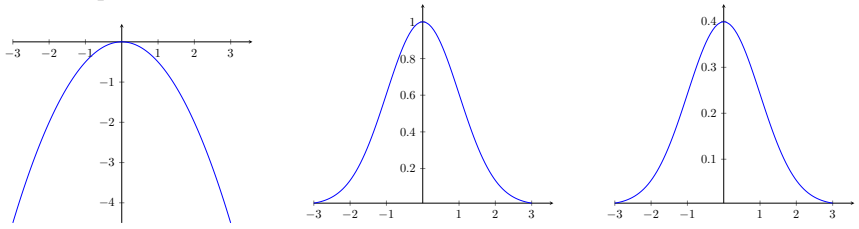
2.2.7 Distribution gaussienne

Cas univarié

La fonction de densité de probabilité la plus utilisée est celle de la distribution normale définie pour $x \in \mathbb{R}$.

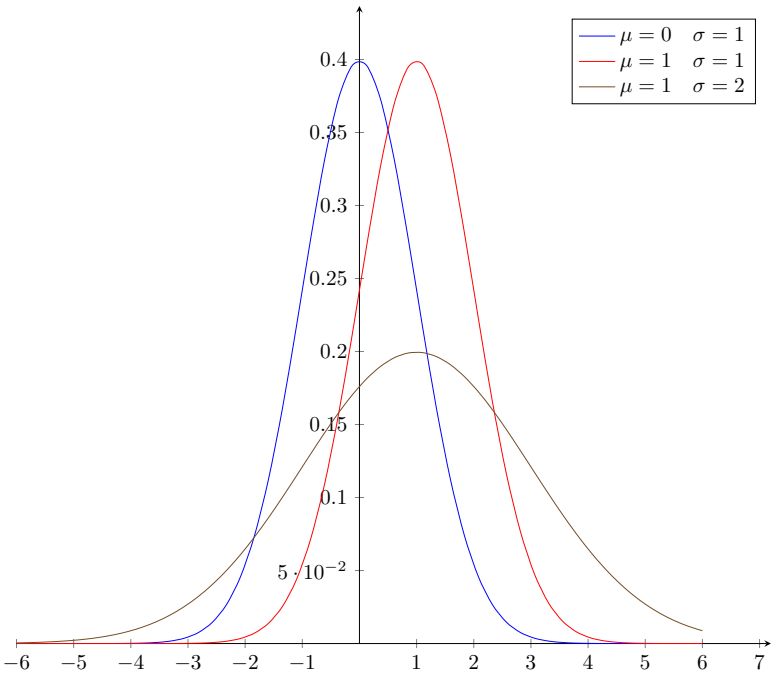
$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big). \tag{2.25}$$

Comme illustré ci-dessous, il s'agit d'une parabole "négative" (graphe de gauche) à laquelle on applique une fonction exponentielle (graphe de droite). On obtient ainsi une fonction positive et symétrique, centrée sur l'axes des y. On normalise cette fonction pour que l'aire sous la groupe soit égale à 1 et obtenir une fonction de densité de probabilité.



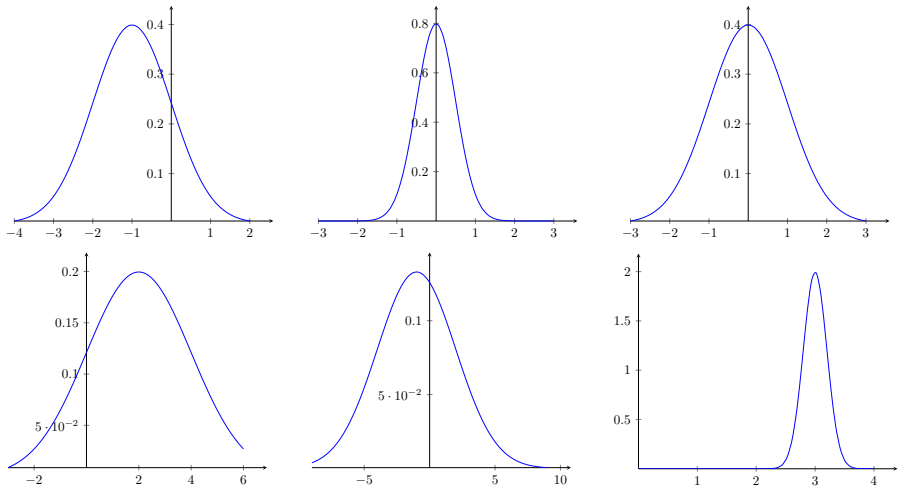
On peut ensuite généraliser cette fonction de la manière à la décaler horizontalement, de manière à ce que sont maximum soit différent de 0. Un décalage de μ vers la droite se fait en remplaçant simplement x par $x - \mu$ dans la fonction. Comme cela ne fait que déplacer la courbe horizontalement, ceci n'a pas d'incidence sur l'aire sous la courbe, qui reste 1.
On peut ensuite modifier vitesse à laquelle $p(x)$ décroit en divisant $x - \mu$ par σ. Pour que l'aire sous la courbe reste toujours 1 il faut alors diviser cette fonction par σ, ce qui donne :

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big). \tag{2.26}$$



Exercice 2.21 Associer les valeurs de paramètre suivants aux fonctions gaussiennes ci-dessous :

$$\mu = -1 \quad \mu = 0 \quad \mu = 2 \quad \mu = 3 \quad \sigma = 0.2 \quad \sigma = 0.5 \quad \sigma = 1 \quad \sigma = 2 \quad \sigma = 3$$



Exercice 2.22 Démontrer que la fonction gaussienne correspond bien à une fonction de densité de probabilité.

Exercice 2.23 En utilisant leur définition ci-dessus, calculer l'espérance et la variance d'une variable aléatoire définie par une loi de probabilité gaussienne de paramètres μ et σ.

Cas multivarié

On peut considérer la distribution conjointe de deux variables aléatoires X_1 et X_2 distribuées selon des distributions gaussiennes de moyennes μ_1 et μ_2 et de variances σ_1^2 et σ_2^2. Cette distribution est données par *[compléter]*

2.2.8 Lois des grands nombres

Les lois des grands nombres permettent d'asseoir le lien entre les modèles théoriques que représentent les variables aléatoires et les observations de la réalité. Elle sont constituées de plusieurs théorèmes qui disent que si on considère n variables aléatoires i.i.d. X_i de loi $p(x)$, les fréquences relatives des réalisations x_i correspondantes tendent vers la distribution $p(X)$ lorsque n tend vers l'infini. De même la moyenne des x_i tend vers $E(X)$.

$$\lim_{n\to\infty} \frac{1}{n} \sum_i x_i \to E(X) \tag{2.27}$$

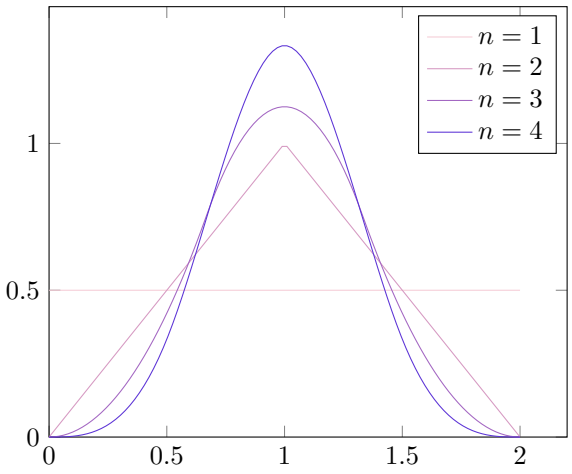
De plus, le *théorème central limite*, nous donne une indication imortante non triviale concernant la distribution de cette moyenne si les X_i ont une variance

σ^2 finie. Il nous dit que si n est suffisamment grand, alors la variable aléatoire définie par la moyenne des X_i se tendra vers une loi gaussienne d'espérance μ et de variance $\frac{\sigma^2}{n}$, où μ est l'espérance des X_i.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \to \mathcal{N}(\mu, \frac{\sigma^2}{n}) \,. \tag{2.28}$$

Ce qui est remarquable dans ce théorème (dont la démonstration dépasse le cadre de ce cours) est que même si les X_i ne sont pas distribuées selon un loi gaussienne, leur moyennne va se rapprocher de plus en plus d'une loi gaussienne. Ainsi, ce dernier théorème indique que la distribution gaussienne a un statut spécial par rapport à d'autres distributions. Il s'agit de la distribution la moins "structurée" ou la moins informative parmi toutes les distributions de même variance.

Exemple 2.10 On considère des variables aléatoires X_i uniformes sur l'intervalle $[0, 2]$. Si on désigne par $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne des n X_i, on peut voir graphiquement ci-dessous que la distribution de Z_n ressemble de plus en plus à une gaussienne lorsque n augmente.



Exercice 2.24 En utilisant les propriétés (démontrées dans les exercices 2.19 et 2.11) que $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ et que $V(kX) = k^2\mathrm{Var}(X)$, démontrer que l'espérance et la variance de la moyenne des variables aléatoire i.i.d. X_i sont celles données par le théorème central limite.

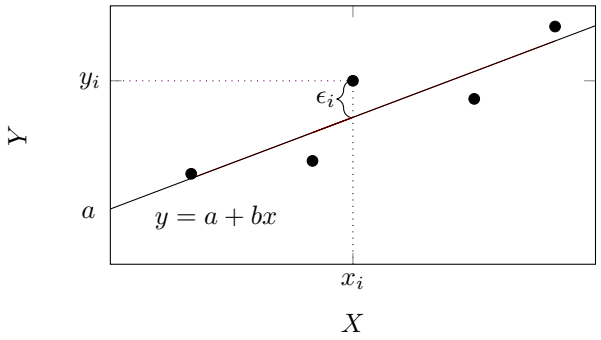


FIGURE 2.1 – Le modèle linéaire

2.3 Modèle linéaire

2.3.1 Modèle univarié

Un des modèles les plus simpes et les plus couramment utilisés est le modèle linéaire. Dans sa version à une dimension, il exprime la relation entre une variable aléatoire expliquée Y et une variables aléatoire explicative X ainsi :

$$Y = a + bX + \epsilon\,, \tag{2.29}$$

où a et b sont des valeurs fixes et ϵ est une variable aléatoire de distribution gaussienne centrée (c'est-à-dire de moyenne 0). Ce modèle fait donc l'hypothèse que X augmente (ou diminue) en moyenne linéairement avec X et que les variables aléatoires X et ϵ sont indépendantes. La variables aléatoire $\epsilon = Y - (a + bX)$ représente l'écart entre Y et $a + bX$ représente la part *inexpliquée* du modèle, qu'on appelle aussi *le bruit* ou *les résidus*.

Ce modèle portant sur la relation entre plusieurs variables aléatoires, on peut le traduire en un modèle portant sur les réalisations de ces variables aléatoires. En ce cas, chaque réalisation produire une valeur x pour X et une valeur y pour Y. Pour les distinguer les unes des autres, on peut les indexer par une variable i, obtenant ainsi une série de paire (x_i, y_i). On peut alors formuler notre modèle ainsi :

$$y_i = a + bx_i + \epsilon_i\,, \tag{2.30}$$

où les ϵ_i sont les réalisations de ϵ. On peut auss mettre tous les x_i, y_i et ϵ_i dans des vecteurs $\vec{x}$, $\vec{y}$ et $\vec{\epsilon}$ et exprimer le modèle ainsi :

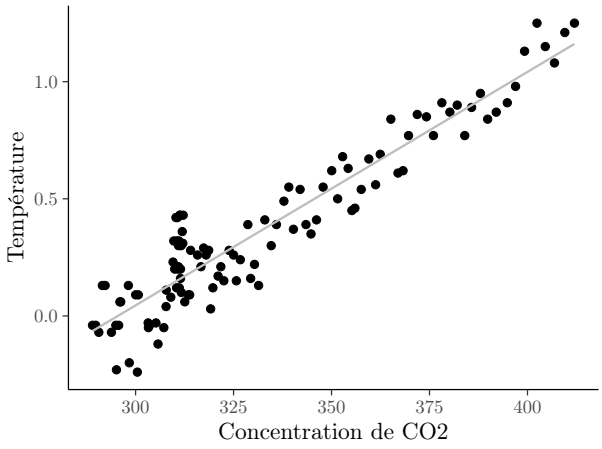
$$\vec{y} = a\vec{1} + b\vec{x} + \vec{\epsilon}, \tag{2.31}$$

où $\vec{1}$ est un vecteur ne contenant que des 1. Au final, par abus de notation, on écrit souvent simplement

$$y = a + bx + \epsilon\,, \tag{2.32}$$

et on laisse le contexte déterminer si on parle des variables aléatoires, des réalisations ou des deux. Ce qui est important à retenir, c'est que les valeurs de x, y et ϵ changent alors que celles de a et b sont fixes. Ces derniers sont des paramètres du modèle.

Exemple 2.11 On souhaite examiner le lien entre le réchauffement climatique et la concentration de CO_2 dans l'atmosphère. Au vu des connaissances scientifiques, on peut se dire que la variables X est la concentration de CO_2 et la variable expliquée Y est la température. Pour obtenir les réalisation de ces variables aléatoires, on prend la température moyenne du globe et la concentration moyenne de CO_2 pour plusieurs années, chaque année correspondant à une réalisation. [1]



Exercice 2.25 On considère le modèle linéaire suivant

$$Y = a + bX + \epsilon\,, \tag{2.33}$$

où le résidu ϵ suit une distribution Gaussienne de paramètres $\mu = 0$ et σ^2. Déterminer la fonction de densité conditionnelle $p(y|x)$ de Y en fonction de X.

2.3.2 Modèle linéaire multivarié

Dans la pratique, un modèle linéaire a souvent plus d'une variable explicative. Si on a deux variables explicatives X et Z, on peut proposer le modèle suivant :

$$Y = a + bX + cZ + \epsilon\,. \tag{2.34}$$

1. Les données sont issues de `https://factsonclimate.org/infographics/concentration-warming-relationship`

Les réalisations produites par ce modèle peuvent alors être représentées dans l'espace à trois dimension par des points distribués autour d' un plan.
[ajouter une illustration]

Si on en a plus et qu'on ne souhaite pas utiliser toutes les lettres de l'alphabet, on peut représenter les différentes variables aléatoires avec les symboles X_1, X_2, X_3,.., X_n. On a alors le modèle suivant

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_n X_n + \epsilon\,, \tag{2.35}$$

qui peut être notée de manière plus compacte ainsi

$$Y = a + \sum_{j=1}^{n} b_j X_j + \epsilon\,, \tag{2.36}$$

ou, en mettant touts les b_j dans un vecteur $\vec{b}$ et les X_i dans un vecteur $\vec{X}$,

$$Y = a + \vec{b} \cdot \vec{X} + \epsilon\,, \tag{2.37}$$

où $\cdot$ représente le produit scalaire usuel.

Si, comme évoqué ci-dessus, on a plusieurs réalisations de chaque variable aléatoire, alors on peut les représenter dans une matrice $\mathbf{X}$, où chaque ligne est une réalisation et chaque colonne une variable aléatoire. Les variables expliquées y_i sont alors représentées par un vecteur $\vec{y}$:

$$\vec{y} = a\vec{1} + \mathbf{X} \cdot \vec{b} + \vec{\epsilon}. \tag{2.38}$$

2.3.3 Modèle linéaire logarithmique

Il arrive souvent que les résidus s'expriment proportionnellement par rapport à la valeur prise par une variable aléatoire. C'est souvent le cas de variables aléatoires continues qui sont par définition positive et qui peuvent couvrir plusieurs ordres de magnitude.

2.4 Modèle logistique

Le modèle logisitique a été développé pour modéliser une variable aléatoire discrète Y binaire, c'est-à-dire pouvant uniquement prendre les valeur 0 ou 1. L'idée de ce modèle est d'exprimer la probabilité conditionnelle de Y en fonction d'une variable explicative X

$$P(Y = 1|X = x) = p(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}\,. \tag{2.39}$$

Si on néglige μ et s pour l'instant, on obtient une courbe *sigmoïde*, c'est-à-dire une courbe en forme de "S"' comprise entre 0 et 1, ce qui justifie qu'on

peut l'interpréter comme une probabilité. Comme pour la gaussienne, le paramètre μ, permet de décaler cette courbe horizontalement et le paramètre s, permet d'étaler ou resserer la courbe horizontalement. On remarque que lorsque $x - \mu$, $p(y|x = \mu) = \frac{1}{2}$, donc μ correspond à la valeur de X pour laquelle $P(Y = 0) = P(Y = 0) = \frac{1}{2}$, c'est à dire que les probabilité des deux valeurs possibles pour Y sont les mêmes.
Une différence importante entre le modèle logistique et le modèle linéaire est que le modèle logistique modélise la loi de probabilité de Y en fonction de X, alors que le modèle linéaire modélise directement la valeur de Y en fonction de X.

Exemple 2.12 Un acheteur souhaite acheter un meuble détenu par une vendeuse potentielle. On peut décrire le montant proposé par l'acheteur par la variable X, et la l'acceptation ou pas de l'offre par la vendeuse par une variable binaire Y qui vaut 1 si la vendeuse accepte, 0 si la vendeuse refuse. On peut modéliser la relation entre l'offre de l'acheteur est son acceptation par le modèle logistique suivant :

$$p(y|x) = \frac{1}{1 + e^{-\frac{x-100}{10}}}\,, \tag{2.40}$$

que l'on peut représenter par le graphique ci-dessous.

Ceci indique que pour un prix de 100, la personne a une chance sur deux d'accepter, et avec un prix de 140, c'est presque sûr qu'elle accepte, alors que pour un prix de 60, c'est presque sûr qu'elle refuse.

Exercice 2.26 Le petit Paul demande à sortir en T-shirt par un matin d'hiver. Ses parents modélisent la probabilité qu'il tombe malade en fonction de la température, par un modèle logistique de paramètre $\mu = 5$ et $s = -2$. Ils sont d'accord que leur fils prenne au maximum un risque de 10% de tomber malade. A partir de quelle température vont-il donner leur autorisation selon le modèle ?

Exercice 2.27 Démontrer que dans cas du modèle logistique définit à l'équation 2.39, on a pour $\mu = 0$,

$$P(Y = 0|X = x) = p(-x) \tag{2.41}$$

Comment ceci peut-il s'interpréter graphiquement ?

Exercice 2.28 Démontrer que le modèle logistique peut être reparamétriser avec les paramètres a et b de la manière suivante :

$$p(y|x) = \frac{1}{1 + e^{-(a+bx)}}\,. \tag{2.42}$$

2.4.1 Modèle logistique multivarié

De la même façon que le modèle linéraire peut être étendu au cas avec plusieurs variables aléatoires explicatives, on peut généraliser le modèle logistique au cas multivarié.

$$p(y|x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{-(a+b_1x_1+b_2x_2+\cdots+b_nx_n)}} \tag{2.43}$$

$$p(y|\vec{x}) = \frac{1}{1 + e^{-(a+\vec{b}\cdot\vec{x})}}\,, \tag{2.44}$$

où $\vec{x}$ et le vecteur composé de toutes les variables aléatoires explicatives du modèle.

2.5 Modèle multi-gaussien

Les deux modèles ci-dessus, le modèle linéaire et le modèle logistique, sont par définition monontones, c'est à dire que Y (ou sa probabilité dans le cas logistique) est soit croissant soit décroissant avec X. Cela limite grandement les phénomènes que ces modèles peuvent représenter, car tous les effets non-monotones ne peuvent pas être capturés par ces modèles. Le modèle mutli-gaussien est plus général dans le sens où la classe des relations qu'il peut représenter est beaucoup plus large.

2.5.1 Mélange de gaussiennes

Il peut arrivé que la distribution d'une variable aléatoire ait plusieurs modes, par exemple si elles sont reparties autour de deux valeurs principales. Dans ce cas, on peut les représenter leur fonction densité de probabilité comme la somme pondérée de deux fonctions gaussiennes.

$$p(x) = \pi_1 g_{\mu_1,\sigma_1}(x) + \pi_2 g_{\mu_2,\sigma_2}(x) \tag{2.45}$$

où

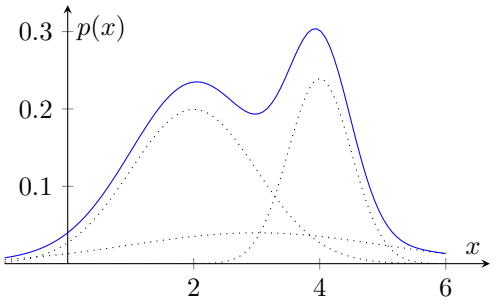
$$g_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big). \tag{2.46}$$

et $\pi_1 + \pi_2 = 1$.
De manière plus générale on peut imaginer une fonction de densité de probabilité $p(x)$ faite d'une somme pondérée de k gaussiennes :

$$p(x) = \sum_{i=1}^{k} \pi_i g_{\mu_i,\sigma_i}(x) \quad \text{avec} \quad \sum_{i=1}^{k} \pi_i = 1\,. \tag{2.47}$$

Cette fonction est souvent appelée un mélange de gaussienne (*gaussian mixture* en anglais). On peut démontrer que toute fonction de densité de probabilité peut être approximée avec une précision arbitraire par un mélange de gaussiennes (avec suffisamment de gaussiennes).

Exemple 2.13 La fonction de densité de probabilité représentée en trait plein ci-dessous est un mélange de gaussiennes composés de trois gaussiennes représentées en pointillés.



Exercice 2.29 Calculer l'espérance d'une variable aléatoire X de fonction de densité de probabilité multigaussienne, de paramètres $\{\pi_i, \mu_i, \sigma_i\}$ avec $i = 1, 2, \ldots k$.

2.6 Solutions des exercices

Exercice 2.4

(a) On a $F(a) = P(X < a) = 0$ et $F(b) = P(X < b) = 1$.

(b) On a :
$F'(x) = \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{P(X<x+h)-P(X<x)}{h} = \lim_{h\to 0} \frac{P(x<X<x+h)}{h} = p(x)$.
La première égalité correspond à la définition de la dérivée, la deuxième de la définition de F, la troisième découle des propriétés des probabilités, et la dernière découle de la définition de la densité de probabilité.

Exercice 2.9 $\frac{1}{\alpha}$

Exercice 2.11 Les calculs sont données pour le cas continu. Les cas discret est analogue, mais un peu plus simple. Soit la variable aléatoire $Y = X + k$. La fonction de densité de probabilité de Y est donnée par $p_Y(y) = p_X(y - k)$ où $p_X(x)$ est la fonction de densité de probabilité de X.

(a) On a :

$$\begin{aligned} E(Y) &= \int_{y\in\Omega_Y} p_Y(y)\cdot y \; dy \quad | \quad y = x + k \\ &= \int_{x\in\Omega_X} p_X(x)\cdot (x+k) \; dx \\ &= \int_{x\in\Omega_X} p_X(x)\cdot x + p_X(x)\cdot k \; dx \\ &= \int_{x\in\Omega_X} p_X(x)\cdot x dx + \int_{x\in\Omega_X} p_X(x)\cdot k \; dx \\ &= E(x) + k\int_{x\in\Omega_X} p_X(x) \; dx \\ &= E(x) + k \end{aligned}$$

(b) On a

$$\begin{aligned} \mathrm{Var}(Y) &= \int_{y\in\Omega_Y} p_Y(y)\cdot \big(y - E(Y)\big)^2 \; dy \quad | \quad y = x + k \\ &= \int_{x\in\Omega_X} p_X(y)\cdot \big(x + k - E(X) - k\big)^2 \; dx \\ &= \int_{x\in\Omega_X} p_X(y)\cdot \big(x - E(X)\big)^2 \; dx \\ &= \mathrm{Var}(X) \end{aligned}$$

(c) Soit la variable aléatoire $Y = kX$. La fonction de densité de probabilité de Y est donnée par $p_Y(y) = \frac{1}{k}p_X(\frac{y}{k})$, où $p_X(x)$ est la fonction de densité

de probabilité de X. On a

$$\begin{aligned} E(Y) &= \int_{y\in\Omega_Y} p_Y(y)\cdot y\ dy \quad | \quad y = kx \\ &= \int_{x\in\Omega_X} \frac{1}{k} p_X(x)\cdot kx\ kdx \\ &= k\int_{x\in\Omega_X} p_X(x)\cdot x\ dx \\ &= kE(X) \end{aligned}$$

(d) On a

$$\begin{aligned} \mathrm{Var}(Y) &= \int_{y\in\Omega_Y} p_Y(y)\cdot \big(y - E(Y)\big)^2\ dy \quad | \quad y = kx \\ &= \int_{x\in\Omega_X} \frac{1}{k} p_X(x)\cdot \big(kx - kE(X)\big)^2 kdx \\ &= \int_{x\in\Omega_X} \frac{1}{k} p_X(x)\cdot k^2\big(x - E(X)\big)^2 kdx \\ &= k^2\int_{x\in\Omega_X} p_X(x)\cdot \big(x - E(X)\big)^2 dx \\ &= k^2\mathrm{Var}(X) \end{aligned}$$

(e) On en déduit que $E(aX+b) = aE(X)+b$ et $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$.

Exercice 2.12 $p(x) = \frac{1}{2} \quad \forall x$

Exercice 2.15

(a)

	$p(X\|Y)$	X: S	X: V
Z	Femme	296/402	106/402
	Homme	146/716	570/716
	Enfant	56/109	53/109

(b)

	$p(Z\|X)$	Z: Femme	Z: Homme	Z: Enfant
X	S	296/498	146/498	56/498
	V	106/729	570/729	53/729

(c)

	$p(Z\|Y)$	Z: Femme	Z: Homme	Z: Enfant
Y	1	144/325	175/325	6/325
	2	93/271	154/271	24/271
	3	165/631	387/631	79/631

Exercice 2.21 De gauche à droite et de haut en bas : $\mu = -1; \sigma = 1 \quad \mu = 0; \sigma = 0.5 \quad \mu = 0; \sigma = 1 \quad \mu = 2; \sigma = 0.2 \quad \mu = -1; \sigma = 3 \quad \mu = 3; \sigma = 0.2$

Exercice 2.25 $P(Y = y|X = x) = P(Y = a+bx+\epsilon) = P(\epsilon = y-a-bx) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y-a-bx)^2}{2\sigma^2}\Big)$.

Exercice 2.26 Il faut résoudre $\frac{1}{1+e^{-\frac{x-\mu}{s}}} = \frac{1}{10} \Leftrightarrow 1+e^{-\frac{x-\mu}{s}} = 10 \Leftrightarrow e^{-\frac{x-\mu}{s}} = 9 \Leftrightarrow -\frac{x-\mu}{s} = \log(9) \Leftrightarrow x - \mu = -s\log(9) \Leftrightarrow x = -s\log(9) + \mu \approx 9.4$ degrés.

Exercice 2.27 Pour $\mu = 0$, on a : $P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = 1 - p(x) = 1 - \frac{1}{1+e^{-\frac{x}{s}}} = \frac{1+e^{-\frac{x}{s}}-1}{1+e^{-\frac{x}{s}}} = \frac{1}{e^{\frac{x}{s}}+1} = p(-x)$. Graphiquement, cela signifie que la fonction $p(x)$ est symétrique par symmétrie centrale autour du point $(0, \frac{1}{2}$ si $\mu = 0$. Si ce n'est pas le cas, ce résultat se généralise avec un centre de symétrie au point $(\mu, \frac{1}{2})$.

Exercice 2.29 $\sum_{i=1}^{k} \pi_i \mu_i$

Chapitre 3

Ajustement

3.1 Introduction

Nous avons vu au chapitre **??** ce que sont les données, et au chapitre **??** ce qu'est un modèle. Dans ce chapitre, nous allons voir comment adapter un modèle aux données. Plus précisément l'ajustement d'un modèle (*model fitting* en anglais) consiste à déterminer ses paramètres en fonction d'un jeu de données. Ainsi, en présence de données $\mathcal{X}$, on va choisir un modèle $\mathcal{M}_{\vec{\theta}}$ dont les paramètres sont regroupés dans un vecteur $\vec{\theta}$. On va ensuite chercher à estimer la valeur de ces paramètres qui correspond le mieux aux données $\mathcal{X}$. Les valeurs de paramètres ainsi obtenues correspondent aux *paramètres estimimé* $\hat{\theta}$.

3.2 Vraisemblance

La vraisemblance (*likelihood* en anglais) $V(\mathcal{X}, \mathcal{M})$ d'un jeu de données $\mathcal{X} = \{x|i = 1, 2, \ldots, n\}$ en fonction d'un modèle $\mathcal{M}$ est la probabilité des données x_i selon le modèle $\mathcal{M}$, c'est-à-dire $P(\mathcal{X}|\mathcal{M})$.

$$V(\mathcal{X}, \mathcal{M}) = P(\mathcal{X}|\mathcal{M}) \tag{3.1}$$

Exemple 3.1 On considère un modèle $\mathcal{M}$, constitué d'une distribution gaussienne $g_{\mu,\sigma}$ et ainsi que n observations x_i. Si on fait l'hypothèse que les données x sont i.i.d, la vraisemblance de ces données selon $\mathcal{M}$ vaut

$$\begin{aligned} V(\mathcal{X}, \mathcal{M}) &= P(x_1, x_2, \ldots x_n|\mathcal{M}) = \prod_{i=1}^{n} P(x_i|\mathcal{M}) = \prod_{i=1}^{n} g_{\mu,\sigma}(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i i-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^{n} e^{-\frac{(x_i i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}, \end{aligned}$$

où la première égalité est justifiée par l'hypothèse que les données x_i sont indépendantes entre elles.

Exercice 3.1 Dans l'exemple 2.3 du jet de deux dés, calculer la vraisemblance des données suivantes correspondant à trois lancers de deux dés : $\mathcal{X} = \{8, 9, 10\}$.

3.3 Principe du maximum de vraisemblance

Une méthode couramment utilisée pour ajuster les paramètres d'un modèle consiste à prendre les paramètres qui maximisent la vraisemblance des données par rapport à ce modèle. Exprimé mathématiquement, on a

$$\hat{\theta} = \underset{\vec{\theta}}{\operatorname{argmax}} V(\mathcal{X}, \mathcal{M}_{\vec{\theta}}) \,. \tag{3.2}$$

Pour trouver la valeur $\hat{\theta}$ qui optimise la vraisemblance $V(\mathcal{X}, \mathcal{M}_{\vec{\theta}})$ des données, on peut simplement trouver le zéro de sa dérivée, c'est-à-dire résoudre

$$\frac{d}{d\vec{\theta}} V(\mathcal{X}, \mathcal{M}_{\vec{\theta}}) = 0 \,. \tag{3.3}$$

Exemple 3.2 : Chute de tartines Julie a l'impression que lorsqu'elle fait tomber sa tartine par terre, celle-ci tombe presque tout le temps avec le côté beurré par contre le sol. Elle modélise la situation avec une variable aléatoire binaire X pouvant prendre les valeurs "Beurre" ou "Pain" selon le côté de la tartine qui touche le sol, et dont la loi de probabilité est la suivante.

x	Beurre	Pain
P(X=x)	y	1-y

Cette loi de probabilité est paramétrisée par la variable y, un nombre compris entre 0 et 1. On l'appelle la *distribution de Bernoulli*, en référence au mathématicien bâlois du $XVII^e$ siècle Jakob Bernouilli (qui a par ailleurs découvert le nombre e et formulé la première version de la loi des grands nombres).
Sur une période de plusieurs jours, Julie compte que sa tartine est tombée 6 fois sur le côté beurre et 4 fois sur le côté pain. La vraisemblance de ces observations $\mathcal{X}$ en fonction du modèle est donnée par

$$V(\mathcal{X}, \mathcal{M}_y) = y^6(1-y)^4. \tag{3.4}$$

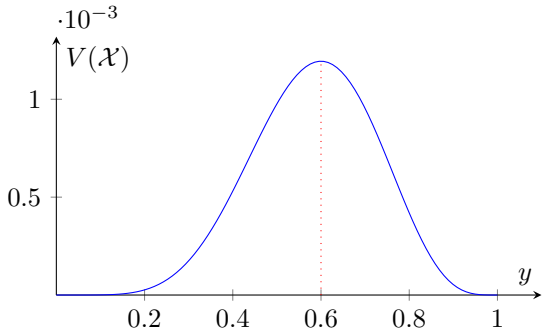
Pour estimer y, on peut utiliser le principe du maximum de vraisemblance

$$\hat{y} = \underset{y}{\operatorname{argmax}} V(\mathcal{X}, \mathcal{M}_y) = \underset{y}{\operatorname{argmax}}\, y^6(1-y)^4. \tag{3.5}$$

On peut calculer $\hat{y}$ en cherchant les zeros de $\frac{d}{dy}V(\mathcal{X}, \mathcal{M}_y)$, c'est-à-dire en résolvant l'équation suivante :

$$\begin{aligned}
& \frac{d}{dy}V(\mathcal{X}, \mathcal{M}_y) = 0 \\
\Leftrightarrow \quad & \frac{d}{dy}y^6(1-y)^4 = 0 \\
\Leftrightarrow \quad & 6y^5(1-y)^4 - 4y^6(1-y)^3 = 0 \\
\Leftrightarrow \quad & y^5(1-y)^3\big(6(1-y) - 4y\big) = 0 \\
\Leftrightarrow \quad & 6(1-y) - 4y = 0 \quad \text{si } y \neq 0 \text{ et } y \neq 1 \\
\Leftrightarrow \quad & 6 - 10y = 0 \\
\Leftrightarrow \quad & 10y = 6 \\
\Leftrightarrow \quad & y = \frac{6}{10}
\end{aligned}$$

On obtient donc sans surprise un seul extrema à $\hat{y} = 0.6$ qui correspond à la proportion de fois où la tartine est tombée sur le côté beurré. On peut s'assurer qu'il s'agit bien d'un maximum et pas d'un minimum car $V(\mathcal{X}, \mathcal{M}_y) \geq 0$ et le minimum 0 est atteint lorsque $y = 0$ ou $y = 1$. On peut visualiser la vraisemblance des donnée en fonction de y avec le graphe suivant :



La vraisemblance est bien maximale pour $y = 0.6$.

Dans les faits, il est souvent plus simple de dériver le logarithme de $V(\mathcal{X}, \mathcal{M}_{\vec{\theta}})$ (la log-vraisemblance, ou *log-likeklihood* en anglais). Comme le log est une fonction strictement croissante, les zéros de la dérivée d'une fonction $f(x)$ sont les mêmes que ceux de la dérivée de $\log(f(x))$ (cf exercice 3.2). De plus, une probabilité (donc de la vraisemblance) étant forcément positive, le logarithme de la vraisemblance sera toujours défini.

Exercice 3.2 Prouver que les zéros de la dérivée de $f(x)$ sont les mêmes que les zéros de la dérivée de $\log(f(x))$, autrement dit

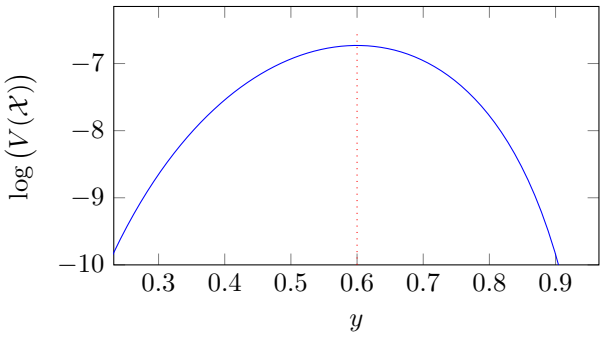
$$f'(x) = 0 \Leftrightarrow \frac{d}{dx}\log\big(f(x)\big) = 0\,, \tag{3.6}$$

où f' représente la dérivée de f.

Exemple 3.2 (suite) On peut arriver au même résultat en cherchant le maximum de la log-vraisemblance $\log\big(V(\mathcal{X}, \mathcal{M}_{\vec{\theta}})\big)$. En reprenant l'équation 3.4, on a :

$$\begin{aligned}
& \frac{d}{dy}\log\big(V(\mathcal{X}, \mathcal{M}_y)\big) = 0 \\
\Leftrightarrow \quad & \frac{d}{dy}\log\big(y^6(1-y)^4\big) = 0 \\
\Leftrightarrow \quad & \frac{d}{dy}\Big(\log(y^6) + \log\big((1-y)^4\big)\Big) = 0 \\
\Leftrightarrow \quad & \frac{d}{dy}\Big(6\log(y) + 4\log(1-y)\Big) = 0 \\
\Leftrightarrow \quad & \frac{6}{y} - \frac{4}{1-y} = 0 \\
\Leftrightarrow \quad & 6(1-y) - 4y = 0 \quad \text{si } y \neq 0 \text{ et } y \neq 1 \\
\Leftrightarrow \quad & 6 - 10y = 0 \\
\Leftrightarrow \quad & 10y = 6 \\
\Leftrightarrow \quad & y = \frac{6}{10}
\end{aligned}$$

Comme on peut s'y attendre, le maximum de la log-vraisemblance est bien au même endroit que le maximum de la vraisemblance, ce qu'on peut également voir en tirant le graphe de la log-vraisemblance, comme ci-dessous.



On remarque que lorsque y tend vers 0 ou 1, la vraisemblance $V(\mathcal{X}, \mathcal{M}_y)$ tend vers 0, et donc la log-vraisemblance $\log\big(V(\mathcal{X}, \mathcal{M}_y)\big)$ tend vers $-\infty$,ce qui explique pourquoi le graphe est "coupé".

Exercice 3.3 Un institut de sondage a interrogé cent personnes qui ont indiqué si elles avaient l'intention de voter pour ou contre un objet de votation, ou si elles sont encore indécises. Ces observations sont modélisées par une modèle $\mathcal{M}$ défini par une variable aléatoire discrète X dont la loi de probabilité est la suivante :

x	Pour	Contre	Indécis
P(X=x)	y	z	1-y-z

Les paramètres de ce modèles sont y et z, deux nombres compris entre 0 et 1. Sur les 100 personnes interrogée, 40 ont indiqué être pour, 35 ont indiqué être contre et 25 sont indécises.

(a) Exprimer la vraisemblance des données récoltées

(b) Exprimet la log-vraisemblance des données récoltées

(c) Appliquer le principe du maximum de vraisemblance pour déterminer les paramètres du model.

Exercice 3.4 On considère une variable aléatoire continue dont la distribution est uniforme entre a et b. Les données $\mathcal{X}$ sont constituées de n valeurs x_i. Utiliser le principe du maximum de vraisemblance pour déterminer les estimations $\hat{a}$ et $\hat{b}$ des paramètres a et b du modèle en fonction de $\mathcal{X}$.
Indice: Il n'est pas nécessaire de dériver la vraisemblance (ni la log-vraisemblance).

Exercice 3.5 Le calcul de la vraisemblance $V(\mathcal{X}, \mathcal{M}_{\vec{\theta}})$ d'un jeu de données $\mathcal{X}$ en fonction d'un modèle $\mathcal{M}_{\vec{\theta}}$ donne $V(\mathcal{X}, \mathcal{M}_{\vec{\theta}}) = 0 \quad \forall \vec{\theta}$.
Comment interpréter ce résultat et est-ce possible d'estimer $\vec{\theta}$ avec le principe du maximum de vraisemblance ?

3.3.1 Ajustement d'une gaussienne

En reprenant l'exemple 3.1, on peut déterminer les paramètres μ et σ d'une distribution gaussienne. La vraisemblance est donnée par

$$V(\mathcal{X}, \mathcal{M}) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} . \tag{3.7}$$

On remarque que cette expression comporant beaucoup de produits et de puissances, la log-vraisemblance sera sans doute plus facile à dériver. Celle-ci est donnée par

$$\log\big(V(\mathcal{X}, \mathcal{M})\big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} . \tag{3.8}$$

On cherche ensuite le zéro de la dérivée par rapport à μ pour obtenir $\hat{\mu}$. La dérivée est donnée par

$$\frac{d}{d\mu}\log\big(V(\mathcal{X},\mathcal{M})\big) = \frac{d}{d\mu}\Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\Big) = \frac{\sum_{i=1}^n (x_i-\mu)}{\sigma^2}$$

Son zéro est donnée par

$$\begin{aligned}
& \frac{\sum_{i=1}^n (x_i-\mu)}{\sigma^2} = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^n (x_i-\mu) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^n x_i - n\mu = 0 \\
\Leftrightarrow \quad & \mu = \frac{1}{n}\sum_{i=1}^n x_i \,,
\end{aligned}$$

ce qui correspond à la moyenne empirique des x_i. Il est aisé de voir qu'il s'agit d'un maximum car $\frac{d}{d\mu}\log\big(V(\mathcal{X},\mathcal{M})\big)$ décroit avec μ.

Pour estimer la variance σ^2, on a

$$\begin{aligned}
\frac{d}{d\sigma^2}\log\big(V(\mathcal{X},\mathcal{M})\big) &= \frac{d}{d\sigma^2}\Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\Big) \\
&= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^2}
\end{aligned}$$

Son zéro est donnée par

$$\begin{aligned}
& -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2(\sigma^2)^2} = 0 \\
\Leftrightarrow \quad & -n\sigma^2 + \sum_{i=1}^n (x_i-\mu)^2 = 0 \\
\Leftrightarrow \quad & \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 \,.
\end{aligned}$$

ce qui correspond à la variance empirique des x_i. De nouveau, il s'agit d'un maximum car $\frac{d}{d\sigma^2}\log\big(V(\mathcal{X},\mathcal{M})\big)$ décroit avec σ^2.
Ainsi, nous venons de montrer que la moyenne et la variance empirique d'un jeu de données correspondent aux estimations des paramètres μ et σ^2 d'une distribution gaussienne selon le principe du maximum de vraisemblance. Pour résumer, on a

1. une distribution théorique imaginaire $\mathcal{M}$ aux paramètres inconnus μ et σ^2 correspondant à son espérance et à sa variance

2. un jeu de données $\mathcal{X}$ qui peut être caractérisé par une moyenne $\bar{x}$ et une variance empirique $\mathrm{Var}(x)$
3. une estimation $\hat{\mu}$, $\hat{\sigma}^2$ correspondant aux paramètres de la distribution imaginaire qui maximise la probabilité des données $\mathcal{X}$ selont la distribution théorique $\mathcal{M}$.

Nous venons de montrer que $\hat{\mu} = \bar{x}$ et $\hat{\sigma}^2 = \mathrm{Var}(x)$.

Exercice 3.6 Appliquer le principe du maximum de vraisemblance pour estimer les paramètres d'une distribution gaussienne à deux dimensions (cas multivarié).

3.3.2 Ajustement d'un modèle logistique

On peut également utiliséer le principe du maximum de vraisemblance pour ajuster les paramètres a et b d'un modèle logistique donné par

$$p(y|x) = \frac{1}{1 + e^{-(a+bx)}} \,. \tag{3.9}$$

Contrairement au cas de la gaussienne, il n'y a pas de solution analytique déterminant les zéros de la dérivée de la vraisemblance (ou de la log-vraisemblance). On a donc recourt à des méthodes numériques qui fournissent de bonne approximation de ces solutions.

Exercice 3.7 Formuler la vraisemblance $V(\mathcal{Y}|\mathcal{X}, \mathcal{M})$ ou les données $\mathcal{Y} = \{y_i\}$ sont binaires, les données $\mathcal{X} = \{x_i\}$ sont continues et le modèle $\mathcal{M}$ est un modèle logistique donné par l'équation (3.9).

3.4 Méthode des moindres carrés

La méthode des moindres carrés consiste à trouver les paramètres d'un modèle qui minimise la somme des carrés des résidus. L'idée est de minimiser la sommme des carrés des erreurs entre la prédiction du modèle (sans bruit) et le résultat observé. La choix de prendre le carrés des erreurs, tient à la fois à des raison pratique de simplification des calculs, mais aussi pour des considérations théoriques liées au fait que, comme on le verra, cela nous ramène aux notions de variances et de covariances.

3.4.1 Ajustement d'un modèle linéaire

On applique la méthode des moindres carrés au un modèle linéaire $\mathcal{M}$ suivant :

$$y = a + bx + \epsilon \tag{3.10}$$

On cherche les paramètres a et b qui minimisent la somme des carrés des résidus ϵ, c'est à dire :

$$(\hat{a}, \hat{b}) = \underset{a,b}{\operatorname{argmin}} \sum_{i}^{n} \epsilon_i^2 = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3.11}$$

A nouveau on cherche les zéros de la dérivée :

$$\begin{cases} \dfrac{d}{da} \displaystyle\sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0 \\ \dfrac{d}{db} \displaystyle\sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0 \end{cases} \Leftrightarrow \begin{cases} -2 \displaystyle\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ -2 \displaystyle\sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} y_i - na - b \sum_{i=1}^{n} x_i = 0 \\ \displaystyle\sum_{i=1}^{n} y_i x_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 = 0 \end{cases} \tag{3.12}$$

On remarque que $\sum_{i=1}^{n} y_i = n\overline{y}$ où est $\overline{y}$ est la moyenne des y_i, et pareil pour les x_i. Pour alléger la notation, on peut ainsi définir

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \quad \overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

On obtient alors pour le system 3.12

$$\begin{cases} n\bar{y} - na - nb\bar{x} = 0 \\ n\overline{xy} - na - nb\overline{x^2} = 0 \end{cases} \Leftrightarrow \begin{cases} a = \bar{y} - b\bar{x} \\ 0 = \overline{xy} - (\bar{y} - b\bar{x})\bar{x} - b\overline{x^2} \end{cases}$$

$$\Leftrightarrow \begin{cases} a = \bar{y} - b\bar{x} \\ 0 = \overline{xy} - \bar{y}\bar{x} + b\bar{x}^2 - b\overline{x^2} \end{cases} \Leftrightarrow \begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} \end{cases} .$$

On a donc comme résultat :

$$\hat{b} = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} \tag{3.13}$$

$$\hat{a} = \bar{y} - b\bar{x} \,. \tag{3.14}$$

Ce qui nous done les paramètres de la droite qui minimise la somme des carrés des résidus. On sait que c'est un minumum car la dérivée calculée ci-dessus croit avec a et avec b.

Exercice 3.8 Monter que l'équation 3.13 peut être reformulée comme un ratio d'une covariance et d'une variance.

Exercice 3.9 Prouver que la méthode des moindres carrés et le principes du maximum de vraisemblance sont équivalents dans la cas du modèle linéaire à résidus i.i.d. gaussiens.

Exercice 3.10 On considére les deux modèles linéaires suivants établissant une relation entre les variables aléatoires continue X et Y.

$$\mathcal{M}_1 : Y = a_1 + b_1 X + \epsilon_1 \quad \text{et} \quad \mathcal{M}_2 : X = a_2 + b_2 Y + \epsilon_2 \,, \tag{3.15}$$

où ϵ_1 et ϵ_2 suivent une distribution gaussienne et centrée. On utilise la méthode des moindres carrés pour estimer les paramètres de ces deux modèles, et on obtient des valeurs $\hat{a}_1$, $\hat{b}_1$, $\hat{a}_2$ et $\hat{b}_2$. On représente les réalisations des X et Y sur un graphique, ainsi que les droites $d_1 : y = \hat{a}_1 + \hat{b}_1 x$ et $d_2 : x = \hat{a}_2 + \hat{b}_2 y$.

(a) Indiquer sur le graphique les valeurs ϵ_1 et ϵ_2.

(b) Les droites d_1 et d_2 sont-elles superposées ?

(c) Proposer un critère d'optimisation qui soit invariant par rapport à une symmétrie ou une rotation des données (x_i, y_i).

3.5 Variables latentes

Il peut arriver des des données soit modélisées comme une variable aléatoire dépendant d'autres variables aléatoires qui ne sont, elles, pas observées. On appelle *variable latente* une variable aléatoire non observée mais qui influence, selon le modèle, une variables aléatoire observée.

Exemple 2.5 (suite) Dans l'exemple du funambule, on peut imaginer, que selon l'épaisseur de la corde utilisée, la probabilité de tomber sera différente et donc la distribution de la durée avant le chute dépendra de la corde. Si le funambule possède deux cordes, on modéliser la corde utilisée par une variable aléatoire binaire Z et le temps du funambule reste sur la corde par une variable aléatoire X dépendante de Z. On aura un paramètre α_0 caractérisant la distribution de X si $Z = 0$ et un paramètre α_1 caractérisant la distribution de X si $Z = 1$. Si on peut observer uniquement X, mais pas Z, alors Z est une variable latente.

Exemple 3.3 Une variable aléatoire X avec une distribution multi-gaussienne peut être considérée comme une variable aléatoire dont la distribution est gaussienne mais dépend d'une variable latente Z qui indique de quelle gaussienne X est issue. Autrement dit, Z peut prendre k valeurs de 1 à k, selon des probabilité $\pi_1, \ldots, \pi_k$ (avec $\sum_i^k \pi_i = 1$). Selon la valeur z prise par Z, alors X sera issue d'une distribution gaussienne de paramètres (μ_z, σ_z).

3.6 Ajustement d'une multi-gaussienne

On pourrait essayer d'utiliser le principe du maximum de vraisemblance pour estimer les paramètres d'une distribution multigaussienne en fonction d'observations $\mathcal{X} = \{x_i\}$, avec $i = 1, \ldots, n$. En reprenant l'équation 2.47 de la distribution multi-gaussienne, on a

$$V(\mathcal{X}, \mathcal{M}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j g_{\mu_j, \sigma_j}(x_i) \tag{3.16}$$

$$\log\big(V(\mathcal{X}, \mathcal{M})\big) = \sum_{i=1}^{n} \log\big(\sum_{j=1}^{k} \pi_j g_{\mu_j, \sigma_j}(x_i)\big) \tag{3.17}$$

Il n'est pas possible de trouver analytiquement (c'est-à-dire exprimer sous forme de formule mathématique) les zéros des dérivées de la vraisemblance ou de la log-vraisemblance (cf exercice 3.11).
Pour contourner cette difficulté, une possibilité est d'appliquer l'algorithme des k moyennes.

Exercice 3.11 Calculer la dérivée de la log-vraisemblance d'un modèle multi-gaussien par rapport au paramètre μ_l d'une des gaussiennes. Qu'est-ce qui empêche d'en déterminer le zéro ?

3.6.1 Algorithme des k moyennes

L'idée de l'algorithme des k moyennes est de séparer le problème de l'estimation des paramètres d'une multi-gaussienne en deux problèmes interdépendants. Le premier problème consiste à associer une des gaussiennes à chaque observation, c'est-à-dire trouver la gaussienne qui aura plus grande chance d'avoir générer cette observation. Pour résoudre ce problème on a besoin de connaître les paramètres des gaussiennes. Le second problème consiste à estimer les paramètres des gaussiennes, sur la base des points qui ont été générer par celles-ci, ce qui revient au problème d'ajustement d'une distribution gaussienne (cf section 3.3.1). Pour pouvoir résoudre ce problème, il faut avoir résolu le premier problème. L'algorithme des k moyennnes propose donc de commencer avec des paramètres aléatoires, puis d'alterner entre le premier et le second problème jusqu'à la convergence vers une solution stable, c'est-à-dire jusqu'à ce que les paramètre n'évolue plus au cours des itérations. Il peut eut être décrit ainsi :

Les k moyennes

1. **Initialisation** Initialiser les μ_j avec le centre de masse des données à laquelle est ajoutée un peu de "bruit", c'est-à-dire une petite valeur aléatoire

ϵ. Initialiséer également les σ_j avec une valeur arbitraire, par exemple 1.

$$\mu_j = \sum_{i=1}^{n} x_i + \epsilon_j \tag{3.18}$$

$$\sigma_j = 1 \tag{3.19}$$

2. **Assignation** Estimer z_i pour chaque donnée x_i, c'est à dire trouver la gaussienne j qui maximise la vraisemblance de x_i.

$$\hat{z}_i = \underset{j}{\operatorname{argmax}}\, V\big(x_i, (\mu_j, \sigma_j)\big) = \underset{j}{\operatorname{argmax}}\, g_{\mu_j,\sigma_j}(x_i) \tag{3.20}$$

3. **Estimation** Estimer les paramètres π_j, μ_j et σ_j de toutes les gaussiennes en fonction des $(\hat{z}_i, x_i)$.

$$n_j = \sum_{\{i|\hat{z}_i=j\}} 1 \quad \text{est le nombres de points associés à la gaussienne } j.$$
$$\pi_j = \frac{n_j}{n}$$
$$\mu_j = \frac{1}{n_j} \sum_{\{i|\hat{z}_i=j\}} x_i$$
$$\sigma_j^2 = \frac{1}{n_j} \sum_{\{i|\hat{z}_i=j\}} (x_i - \mu_j)^2$$

4. **Arrêt** S'arrêter si les $\hat{z}_i$ sont restés identiques à l'itération précédente, sinon recommencer en 2.

On peut montrer (cf exercice 3.12) que cet algorithme converge vers une solution, mais que celle-ci peut être différente selon l'initialisation du début. L'algorithme n'est donc pas garanti de converger vers la meilleure solution. Dans une version simplifiée et initiale, cet algorithme n'estime pas la variance des gaussiennes qui est considérée comme identique et constante pour toutes les gaussiennes (d'où le nom des "k moyennes", plutôt que "k moyennes et variances").

Exercice 3.12 Prouver que l'algorithme des k moyennes converge toujours en montrant que $P\big(\{X_i = x_i, Z_i = \hat{z}_i\}|\{\pi_j, \mu_j, \sigma_j\}\big)$ augmente aussi bien lors de la phase d'assignation que la phase d'estimation.
Indice: Utiliser la formule des probabilités composées donnée par l'équation 2.18.

3.6.2 Algorithme EM

L'algorithme EM ou de l'espérance maximale (*expectation maximization* en anglais) permet également d'estimer les paramètres d'une multigaussienne, mais il est plus général et peut donc s'appliquer à d'autres modèles impliquant des

variables latentes.
Cet algorithme utilise à nouveau le principe du maximum de vraisemblance. Toutefois, comme la variable latente Z n'est pas observée, on ne peut pas calculer sa vraisemblance car elle dépend des valeurs prises par Z. Mais, si on connait la distribution de Z, alors on peut calculer la vraisemblance marginale de X par rapport à Z. Dans le cas où Z est une variable aléatoire discrète, cela donne

$$\begin{aligned} V(\mathcal{X}, \mathcal{M}) = P(\mathcal{X}|\mathcal{M}) &= \sum_{z=\Omega_Z} P(\{X_i = x_i, Z_i = z\}|\mathcal{M}) \\ &= \sum_{z=\Omega_Z} P(\{X_i|Z_i = z\}, \mathcal{M}) \cdot P(\{Z_i = z\}|\mathcal{M}), \end{aligned} \tag{3.21}$$

où l'accolade dans $P(\{X_i = x_i\})$ etc. indique qu'il s'agit de la probabilité de l'ensemble des données x_i. L'idée de l'algorithme EM, est d'estimer $P(Z_i = z|\mathcal{M})$ en calculant $P(Z_i = z|X_i, \mathcal{M})$, puis d'estimer les paramêtres qui maximisent la vraisemblance ainsi estimée. On obtient ainsi de nouveaux paramètres du modèle qui permettent une meilleure estimation de $P(Z_i = z|\mathcal{M})$ et ainsi de suite.

Appliqué au cas de la distribution multi-gaussienne, cela donne :

EM appliqué à une distribution multigaussienne

1. **Initialisation** Initialiser les μ_j avec le centre de masse des données à laquelle est ajoutée un peu de "bruit", c'est-à-dire une petite valeur aléatoire ϵ. Initialiséer également les σ_j avec une valeur arbitraire, par exemple 1, ainsi que les π_j à $\frac{1}{k}$.
2. **Estimation de la distribution conditionnelle** Estimer $p_{ij} = P(Z_i = j|X_i, \mathcal{M})$ qui est la distribution conditionnelle de Z par rapport à X. Cela représente la probabilité que le point x_i ait été généré par la gaussienne j.

$$p_{ij} = P(Z_i = j|X_i, \mathcal{M}) = \frac{\pi_j g_{\mu_j,\sigma_j}(x_i)}{\sum_{j=1}^{k} \pi_j g_{\mu_j,\sigma_j}(x_i)} \tag{3.22}$$

3. **Maximisation de la vraisemblance ainsi estimée** Estimer les paramètres π_j, μ_j et σ_j de toutes les gaussiennes en fonction des p_{ij}.

$$\pi_j = \frac{\sum_{i=1}^{n} p_{ij}}{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ij}} \tag{3.23}$$

$$\mu_j = \frac{\sum_{i=1}^{n} p_{ij} x_i}{\sum_{i=1}^{n} p_{ij}} \tag{3.24}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{n} p_{ij}(x_i - \mu_j)^2}{\sum_{i=1}^{n} p_{ij}} \tag{3.25}$$

4. **Arrêt** S'arrêter si les p_{ij} sont restés (presque) identiques à l'itération précédente, sinon on recommencer en 2.

En résumé, cet algorithme est très similaire à l'algorithme des k-moyennes, mais au lieu d'attribuer une gaussienne à chaque point (en estimant z_i), il attribue à chaque point une distribution de probabilité d'appartenir à chacune des gaussienne, p_{ij}. Ainsi on pourra avoir des points qui sont attribué à moitié à une gaussienne et à moitée à une autre, ce qui n'est pas possible dans avec les k moyennes.

3.7 Biais et variance d'estimation

Il est utile de savoir si une méthode d'estimation d'un paramètre d'un modèle va avoir tendance à sur-estimer ou sous-estimer la valeur de ce paramètre. C'est ce qu'on appelle les *biais d'estimation*. De manière générale, une méthode *non biaisée*, c'est-à-dire qui n'a pas tendance à sur-estimer ou sous-estimer la valeur du paramètre sera préférable à une méthode biaisée.
Si X est une variable aléatoire dont la loi de probabilité est paramétrisée par un paramètre θ, et $\mathcal{X} = \{x_i\}$ sont des réalisations de cette variable aléatoire, alors on peut appeler $t(X_1, \ldots, X_n)$ une fonction qui permet d'estimer θ en fonction des données $x_1, \ldots, x_n$. La valeur $\hat{\theta}$ prise par cette fonction peut être vue comme une réalisation d'une variable aléatoire T définie à partir des variables aléatoires X_i. Le *biais* de l'estimateur t est alors définit comme

$$B(t) = E\big(T(X_i) - \theta\big)\,, \tag{3.26}$$

où $E()$ est l'espérance. La variance de l'estimateur est donnée par

$$B(t) = E\big(T(X_i) - \theta\big)^2\,. \tag{3.27}$$

Exemple 3.2 (suite) Dans l'exemple de la tartine, on a vu que l'estimateur donné par le maximum de vraisemblances est donné par la proportion de tartines tombées côté "Beurre".

$$t(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} u(x_i) \quad \text{où} \quad u(x) = \begin{cases} 1 & \text{si } x = \text{Beurre} \\ 0 & \text{si } x = \text{Pain} \end{cases} \tag{3.28}$$

Le biais des cet estimateur est donnée par

$$\begin{aligned} B(t) &= E\big(T(X_i) - y\big) = E\big(\frac{1}{n}\sum_{i=1}^{n} u(X_i) - y\big) = \frac{1}{n}\sum_{i=1}^{n} E\big(u(X_i)\big) - E(y) \\ &= \frac{1}{n}\sum_{i=1}^{n} \big(y \cdot 1 + (1-y)\cdot 0\big) - y = \frac{1}{n} ny - y = y - y = 0 \end{aligned}$$

Cet estimateur est donc non-baisé, c'est-à-dire qu'il n'aura pas tendance à sur-estimer ou sous-estimer le paramètre y du modèle.

Exercice 3.13 On considère une variable aléatoire X avec une loi de probabilité uniforme entre 0 et b (avec $b > 0$) ainsi que n réalisation $x_i \in \mathcal{X}$ de X. Déterminer si les estimateurs suivants de b sont biaisés ou pas. Justifier sa réponse.

(a) $t(\mathcal{X}) = \max_i x_i$

(b) $t(\mathcal{X}) = \frac{2}{n} \sum_{i=1}^{n} x_i$

Exercice 3.14 Une variable aléatoire X a une espérance μ et une variance σ^2 qu'on cherche à estimer sur la base de n réalisation x_i, avec $i = 1, \ldots, n$.

(a) Prouver que l'estimateur de la variance donnée par

$$S = \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \quad \text{avec} \quad \bar{x} = \frac{1}{n} \sum_{i}^{n} x_i \tag{3.29}$$

est biaisé.
Indice : Montrer que $E(S) \neq \sigma^2$ en continuant le développement suivant : $E(S) = E\big(\frac{1}{n} \sum_i^n (x_i - \bar{x})^2\big) = E\Big(\frac{1}{n} \sum_i^n \big((x_i - \mu) - (\bar{x} - \mu)\big)^2\Big)$.

(b) En déduire un estimateur non-biaisé de σ^2

3.8 Pré-traitement

3.9 Solutions des exercices

Exercice 3.1 $\frac{5}{3888}$

Exercice 3.2 On a : $\frac{d}{dx}\log(f(x)) = 0 \Leftrightarrow \frac{f'(x)}{f(x)} = 0 \Leftrightarrow f'(x) = 0$

Exercice 3.3

(a) $y^{40}z^{35}(1-y-z)^{25}$

(b) $40\log(y) + 35\log(z) + 25\log(1-y-z)$

(c) On cherche le zero de la dérivée de la log-vraisemblance :

$$\begin{cases} \frac{d}{dy}\Big(40\log(y) + 35\log(z) + 25\log(1-y-z)\Big) = 0 \\ \frac{d}{dz}\Big(40\log(y) + 35\log(z) + 25\log(1-y-z)\Big) = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{40}{y} - \frac{25}{1-y-z} = 0 \\ \frac{35}{z} - \frac{25}{1-y-z} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \frac{40}{y} - \frac{35}{z} = 0 \\ \frac{40}{y} - \frac{25}{1-y-\frac{7}{8}y} = 0 \end{cases} \Leftrightarrow \begin{cases} z = \frac{7}{8}y \\ \frac{8}{y} = \frac{5}{1-\frac{15}{8}y} \end{cases} \Leftrightarrow \begin{cases} z = \frac{7}{8}y \\ 8 - 15y = 5y \end{cases} \Leftrightarrow \begin{cases} y = \frac{4}{10} = 0.4 \\ z = \frac{7}{20} = 0.35 \end{cases}$$

On obtient donc bien $y = 0.4$ et $z = 0.35$ comme escompté.

Exercice 3.5 Cela signifie que le modèle ne peux pas expliquer les données observées donc on ne peut pas appliquer le principe du maximum de vraisemblance. Cela peut être dû au fait que le modèle n'est pas adéquat ou que les observations contiennent des erreurs (ou les deux).

Exercice 3.8 $\hat{b} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$

Exercice 3.9 Dans le cas du modèle linéaire (standard), la vraisemblances est donnée par

$$V(\mathcal{Y},\mathcal{M}) = P(Y_1 = y_1,\dots,Y_n = y_n|a,b,\sigma^2 x_1,\dots x_n) = \prod_{i=1}^{n} P(Y_i = y_i|a,b,X_i = x_i)$$
$$= \prod_{i=1}^{n} P(\epsilon_i = y_i - a - bx_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-a-bx_i)^2}{2\sigma^2}}$$

La log-vraisemblance est donnée par :

$$\log\big(V(\mathcal{Y},\mathcal{M})\big) = \log\Big(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-a-bx_i)^2}{2\sigma^2}}\Big) = -\frac{1}{2}\sum_{i=1}^{n}\big(\log(2\pi\sigma^2) + \frac{(y_i-a-bx_i)^2}{\sigma^2}\big)$$
$$= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-a-bx_i)^2}{\sigma^2}$$

Trouver les paramètres a et b qui maximisent cette valeur revient donc à minimiser $\sum_{i=1}^{n}(y_i - a - bx_i)^2$, autrement dit la somme des carrés des résidus, ce qui correspond à la méthode des moindres carrés.

Exercice 3.11 On arrive à $\sum_i \Big(\sum_j \pi_j g_{\mu_j,\sigma_j}(x_i)\Big)^{-1} \frac{x_i-\mu_l}{\sqrt{2\pi\sigma_l^2}} e^{-\frac{(x_i-\mu_l)^2}{2\sigma_l^2}} = 0$ d'où on ne parvient pas à isoler μ_l car c'est de la forme $\sum_i (x_i-\mu)e^{(x_i-\mu)^2} = 0$.

Exercice 3.13

(a) Oui c'est biaisé car $\max_i x_i < b$. Donc $E(\max_i x_i) < b$ et ne vaut donc pas b.

(b) Non, ce n'est pas biaisé car $E(\frac{2}{n}\sum_{i=1}^n x_i) = \frac{2}{n}E(\sum_{i=1}^n x_i) = \frac{2}{n}\sum_{i=1}^n E(x_i) = \frac{2}{n} n \frac{b}{2} = b$.

Exercice 3.14

(a) On a

$$\begin{aligned}
E(S) &= E\big(\frac{1}{n}\sum_i^n (x_i-\bar{x})^2\big) = E\Big(\frac{1}{n}\sum_i^n \big((x_i-\mu)-(\bar{x}-\mu)\big)^2\Big) \\
&= \frac{1}{n}E\Big(\sum_i^n \big((x_i-\mu)^2 - 2(\bar{x}-\mu)(x_i-\mu) + (\bar{x}-\mu)^2\big)\Big) \\
&= \frac{1}{n}E\Big(\sum_i^n (x_i-\mu)^2 - 2\sum_i^n (\bar{x}-\mu)(x_i-\mu) + \sum_i^n (\bar{x}-\mu)^2\Big) \\
&= \frac{1}{n}E\Big(\sum_i^n (x_i-\mu)^2 - 2(\bar{x}-\mu)\sum_i^n (x_i-\mu) + n(\bar{x}-\mu)^2\Big) \\
&= \frac{1}{n}\sum_i^n E\big((x_i-\mu)^2\big) - 2E\big((\bar{x}-\mu)(\bar{x}-\mu)\big) + E\big((\bar{x}-\mu)^2\big) \\
&= \sigma^2 - E\big((\bar{x}-\mu)^2\big) = \sigma^2 - \mathrm{Var}(\bar{x}) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}$$

(b) Si $E(S) = \frac{n-1}{n}\sigma^2$, alors $E(\frac{n}{n-1}S) = \sigma^2$. Donc $\frac{n}{n-1}S = \frac{1}{n-1}\sum_i^n (x_i-\bar{x})^2$ est un estimateur non biaisé de σ^2.

Chapitre 4

Inférence

4.1 Echantillon et population

Il est souvent impossible ou trop couteux de récolter des données concernant l'entier d'une population. Dans ce cas, on se contente de récolter des données d'un petit sous-ensemble de la population, qu'on appelle un *échantillon*, et on espère que les données correspondant à cet échantillon nous donneront des information sur l'ensemble de la population. Les méthodes consistant à généraliser des informations concernant un échantillon à la population dont il est issu, sont appelées les méthodes *d'inférence*.

Exemple 4.1 : Intentions de vote Un institut de sondage souhaite connaître les intentions de vote de la population suisse. Il ne peut toutefois pas demander leur avis à tous les Suisses, et va donc choisir un échantillon de citoyens et citoyennes suisses et analyser les données issues de cet échantillon. Comme ce n'est pas les intentions de vote de cet échantillon en particulier qui intéresse l'institut de sondage, mais celles de l'entier de la population suisse, des méthodes d'inférences seront utilisées pour tenter de généraliser les observations faites sur cet échantillon à l'ensemble de la population.

Il arrive parfois qu'on a un échantillon, mais qu'on ne sait pas vraiment quelle est la population qui nous intéresse. C'est souvent le cas dans la recherche biomédicale, où des personnes sont recrutées pour des études, mais il est difficile de savoir à quelle population les résultats obtenus sont généralisables.

Exemple 4.2 Les essais cliniques en vue de tester l'efficacité ou la dangerosité d'un médicament ont historiquement souvent exclus les femmes afin de limiter la variabilité due aux cycles hormonaux, ou les effets sur une éventuelle grossesse en cours de traitement. Les résultats obtenus ont ensuite été généralisés aux hommes et aux femmes, ce qui peux se révéler problématique si les femmes et

les hommes réagissent différemment à ces médicaments. [**?**] Le même problème se pose également concernant les personnes issues de minorités éthniques.

Dans une approche rigoureuse, la population aussi bien que l'échantillon devraient être bien définis, et il faudrait s'assurer que l'échantillon obtenu est *représentatif* de la population.

4.1.1 Représentativité

La représentativité d'un échantillon décrit à quel point l'échantillon est semblable à la population. Si c'est le cas, il sera justifié de généraliser les observations faites sur l'échantillon à la population toute entière. Sinon, toute généralisation sera potentiellement abusive. Il est souvent difficile de s'assurer qu'un échantillon est représentatif d'une population, d'autant plus qu'on ne connait pas forcément à priori quelles sont les variables pertinentes. Parfois, on connait certaines données sur l'ensemble de la population ce qui permet d'évaluer la représentativité d'un échantillon.

Exemple 4.1 (suite) L'institut de sondage connaît les caractéristiques de la population suisse et a une certaines idées *a priori* des facteurs qui peuvent influencer sur les intentions de vote, tels que l'âge, le genre, le niveau de revenu, le canton de résidence, ou si la personne habite à la campagne ou en ville. Il pourra donc choisir son échantillon de manière à ce que pour ces facteurs-là leur distributions empiriques soient similaires en l'échantillon et la population. Par exemple, si l'échantillon ne contient qu'une minorité de personnes suisse-allemandes alors que celles-ci forment la grande majorité de la population suisse, on pourra conclure que l'échantillon n'est pas représentatif de la population pour ce qui est de ce facteur, et donc potentiellement également pour les intentions de vote.

Dans certaines situations, on ne connait pas suffisamment bien la population pour avoir une idée de ses charactéristiques et de la représentativité d'un échantillon. Une manière d'avoir de bonnes chances d'obtenir un échantillon représentatif est de sélectionner des individus aléatoirement dans la population, en donnant à chacun la même probabilité d'être sélectionné. Ainsi, si le processus de sélection n'est pas influencé par les caractéristiques des individus, et que les nombres d'individus est suffisant pour refléter la diversité de la population, l'échantillon sera probablement représentatif de la population. Ce n'est toutefois pas toujours si simple de sélectionner un échantillon sans être influencé, même indirectement, par les caractéristiques des individus.

4.1.2 Bias d'échantillonage

Selon la manière dont on sélectionne l'échantillon, on peux induire un biais d'échantillonage, c'est-à-dire que certaines caractéritiques des individus sélec-

tionner seront sur- ou sous-représentées dans notre échantillon par rapport à la population.

Exemple 4.1 (suite) Si l'institut de sondage recrute les individus de son échantillon mettant une annonce sur les réseaux sociaux, il faut s'attendre à ce que les usagers des réseaux sociaux soient sur-représentés dans cet échantillon. A l'inverses, les personnes agées, qui utilisent moins les réseaux sociaux seront sous-représentées.

4.2 Test d'hypothèse

4.2.1 Statistique

Une statistique est une (ou plusieurs) valeurs qui peut être calculée à partir d'un échantillon. Plusieurs statistiques ont déjà été vues dans les chapitres précédents, par exemple la moyenne et la variance empiriques, l'étendue, ou la vraisemblance.

4.2.2 Principe

Le principe du test d'hypothèse consiste d'abord à formuler une hypothèse sur la valeur d'un ou plusieurs paramètres d'un modèle. On utilise ensuite les données pour déterminer à quel point celles-ci sont compatibles avec notre hypothèse. Cette hypothèse, généralement formulé sous la forme d'une égalité est appelée *l'hypothèse nulle* et est symbolisée par $\mathcal{H}_0$. Elle est mise en regard d'une *hypothèse alternative*, souvent notée $\mathcal{H}_1$.

Exemple 3.2 (suite) Dans l'exemple des tartines, on a représenté la probabilité que le la tartine se retrouve face beurrée contre le sol par le paramètre y. Julie a observé que la tartine est tombée quatre fois sur le côté pain est six fois sur le côté beurre. Elle souhaite savoir si sa tartine a vraiment plus de chances de tomber sur le côté beurré (par exemple à cause du poids du beurre) ou si c'est juste une question de malchance. Dans ce dernier cas, elle peut s'attendre à avoir une tartine qui tombe sur le côté beurré à peu prêt la moitié des cas. A contrario, si la tartine a vraiment plus de chances de tomber sur le côté beurré, alors elle peut s'attendre à avoir également dans le futur davantage de beurre sur le parquet. Elle va donc reprendre le modèle de la distribution de Bernoulli et formuler l'hypothèse nulle qui correspond au cas où la tartine a autant de chance de tomber sur chacun des deux côté, autrement dit $\mathcal{H}_0 : y = \frac{1}{2}$.
Concernant l'hypothèse alternative, si Julie suspecte que la tartine a plus de chance de tomber sur le côté beurré, elle pourra être formulée par $\mathcal{H}_1 : y > \frac{1}{2}$.

Une fois les hypothèse $\mathcal{H}_0$ et $\mathcal{H}_1$ formulées, on choisit une statistique T dont il est possible de calculer la distribution sous l'hypothèse $\mathcal{H}_0$. On appelle cette distribution, *la distribution nulle*. On calcule cette statistique pour notre échantillon et détermine sa probabilité selon l'hypothèse $\mathcal{H}_0$. Si cette probabilité est faible, c'est sans doute que l'hypothèse est fausse, alors que si elle est élevée, il est possible que l'hypothèse soit vraie.

Exemple 3.2 (suite) En reprenant l'exemple des tartines, une statistique qui semble appropriées est le nombre de fois où la tartines est tombée sur le côté beurre, sur les 10 chutes. Cette statistique peut être considérée comme une variable aléatoire discrète, nommée T. Sous $\mathcal{H}_0$, on peut calculer la distribution nulle $p(T = t|\mathcal{H}_0)$. Comme $y = \frac{1}{2} = 1 - y$, et que toute les observations ont supposées être i.i.d, toutes les séquences de 10 chutes observables ont la même probabilité donnée par $\frac{1}{2})^{10} = \frac{1}{1024}$. Pour calculer $p(T = t|\mathcal{H}_0)$, il faut calculer la proportion des séquences possible avec t chutes côté beurre. Cette proportion est donnée par le nombre de possibilités d'avoir t "Beurre" et $10-t$ "Pain" dans une séquence de 10 chutes, divisé par le nombre de séquences possibles :

$$p(T = t|\mathcal{H}_0) = \frac{1}{2^{10}} \binom{10}{t} = \frac{10!}{t!(10-t)!2^{10}} \,. \tag{4.1}$$

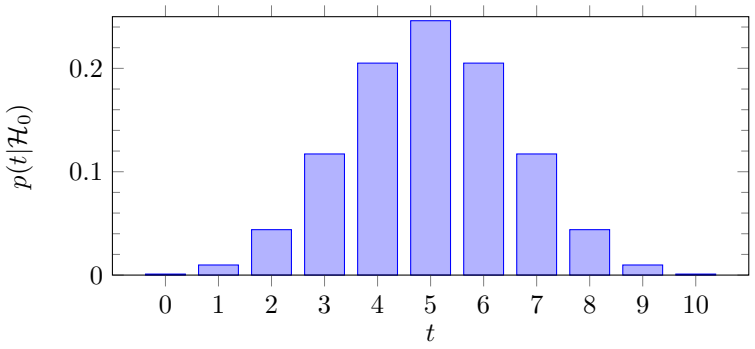
En faisant les calculs, on obtient la distribution nulle suivante.

t	0	1	2	3	4	5	6	7	8	9	10
$p(t\|\mathcal{H}_0)$	$\frac{1}{1024}$	$\frac{10}{1024}$	$\frac{45}{1024}$	$\frac{120}{1024}$	$\frac{210}{1024}$	$\frac{252}{1024}$	$\frac{210}{1024}$	$\frac{120}{1024}$	$\frac{45}{1024}$	$\frac{10}{1024}$	$\frac{1}{1024}$

En effet, parmi les 1024 séquences possible de 10 observations, une seule contient 0 fois "Beurre", il s'agit de la séquene contenant 10 fois "Pain". De même, il y en a 10 qui contiennent exactement une fois "Beurre", 45 qui contiennent exactement 2 fois "Beurre", etc. On peut vérifier qu'il s'agit bien d'une distribution car

$$\sum_{t=0}^{10} p(T = t|\mathcal{H}_0) = 1 \,. \tag{4.2}$$

L'histogramme correspondant à cette distribution peut être représenté ainsi :



Dans certains cas, la statistique choisie peut être l'estimation du paramètre du modèle. Dans l'exemple ci-dessus, on aurait pu choisir comme statistique, non pas le nombres absolu de fois où la tartine est tombée sur le côté beurré, mais la proportion de fois où cela arrive, qui est une estimation du paramètre y selon le principe de maximisation de la vraisemblance. On aurait alors obtenu $p(T' = t) = p(T = 10t)$, où T' serait la nouvelle statistique.

4.2.3 P-valeur

Une fois la distribution nulle obtenue, on peut regarder où se situent nos observations par rapport à cette distribution nulle. Si ces observations sont très improbables selon cette distribution nulle, il y a de bonnes chances que nos hypothèses, en particulier notre hypothèse nulle $\mathcal{H}_0$, soit fausse. On définit la *p-valeur* d'un test, par la probabilité, selon la distribution nulle, d'obtenir une statistique égale ou pire que celle observée. Par "pire", on entends, "plus en accord avec l'hypothèse alternative $\mathcal{H}_1$". Si cette p-valeur, est en dessous d'un seuil décidé intialement appelé *seuil de significativité*, cela signifie que les observations sont peu en accord avec l'hypothèse nulle, et l'hypothèse $\mathcal{H}_0$ est rejetée. Autrement dit, les données sont peut compatibles avec $\mathcal{H}_0$, et favorisent plutôt $\mathcal{H}_1$. Mais si la p-valeur est en dessus de ce seuil, ce n'est pas pour autant que l'on peut accepter $\mathcal{H}_0$. On peut juste dire que les données ne contredisent pas l'hypothèse nulle. Cela peut être dû au fait que l'hypothèse nulle est vraie, mais aussi au fait, très courant, qu'il manque de données pour tirer une conclusion. On ne peut donc rien en déduire.
Le seuil de significativité est choisi par l'analyste selon la confiance dans les résultats souhaitée. On prend souvent 5% ou 1%, mais on peut également prendre des seuils plus faibles selon la quantité de données récoltées et la procédure exprémentale globale.

Exemple 3.2 (suite) Dans l'exemple des chutes de tartines, la probabilité d'observer 6 chutes côté "Beurre" est données par $p(T = 6|\mathcal{H}_0) = \frac{210}{1024}$. Les cas qui sont "pires" sont donnés par $t > 6$. La p-valeur est donc donnée par

$$p(T \geq 6|\mathcal{H}_0) = \sum_{t=6}^{10} p(T = t|\mathcal{H}_0) = \frac{210 + 120 + 45 + 10 + 1}{1024} = \frac{386}{1024} \approx 0.38$$

Ceci indique qui si la tartine à la même probabilité de tomber sur chacune des faces, on a 38% de chances que sur 10 chutes, elle tombe 6 fois ou plus sur le côté "Beurre". Cette probabilité est assez importante, donc les observations sont entièrement compatibles avec notre hypothèse nulle, qui ne devrait de fait, pas être rejetée. Ainsi, on ne peut pas savoir, avec ces observations, si la tartine a tendance à tomber plus souvent du côté beurré. Mais peut-être qu'avec plus d'observations, il serait possible de rejeter notre hypothèse nulle, ou peut-être pas...

Si la p-valeur obtenue est en dessous du seuil de significativité, cela peut être dû au fait que l'hypothèse nulle est fausse, mais cela peut aussi être dû à la malchance. En effet, avec un seuil à 5%, cela signifie que même si $\mathcal{H}_0$ est vraie, dans 5% des cas, on observera des statistiques similaires ou pires que celles que nous avons observées. Autrement dit, on rejetera l'hypothèse nulle à tort dans 5% des cas. Il se peut donc très bien que l'on tombe sur un de ces cas "malchanceux".
Par ailleurs, il se peut aussi que la p-valeur soit faible parce qu'une des hypothèse du modèles, autre que l'hypothèse nulle, n'est pas vérifiée. Ca peut être le cas par exemple si les observations ne sont pas vraiment indépendantes.

Exemple 4.1 (suite) L'institut de sondage d'opinion politique veut effectuer un test d'hypothèse sur la différence d'opinon politique selon l'age. Pour gagner du temps, un employer décide d'interroger tous les membres d'un même foyer plutôt qu'une seule personne par foyer. Comme les opinions politiques sont souvent partagées au sein d'un foyer, les observations ne seront pas indépendantes entre elles. Si le test d'hypothèse postule l'indpendance des observations, alors on pourra observer une p-valeur significative sans que l'hypothèse nulle ne soit fausse pour autant.

Exercice 4.1 - Test binomial

(a) Généraliser l'exemple 3.2 en déterminant la distribution nulle pour le cas où Julie a observé n chutes de tartines.

(b) Généraliser l'exemple 3.2 en déterminant la distribution nulle pour le cas où Julie a observé n chutes de tartines et souhaite tester l'hypothèse nulle $\mathcal{H}_0 : y = a$. Ce test s'appelle le *test binomial*.

Exercice 4.2 Parmi les situations suivantes, indiquer pour lesquelles un test binomial peut être utile. Si c'est le cas, indiquer ce que pourraient être $\mathcal{H}_0$ et $\mathcal{H}_1$, sinon justifier pourquoi.

(a) Les jeunes Alya et Thomas sont en rivalité aux échecs et jouent beaucoup de parties, gagnées tantôt par Alya et tantôt par Thomas. On souhaite savoir si l'un des enfants est meilleur que l'autre.

(b) On souhaite vérifier qu'un dé à jouer est bien équilibré et a autant de chances de tomber sur chacune des faces.

(c) On souhaite savoir si le taux d'échec à un examen est supérieur à 30%.

(d) On souhaite déterminer si un processus de fabrication d'une pièce mécanique produit moins de 1% de pièces défectueuses.

4.2.4 Autres tests paramétriques

L'exemple 3.2 illustre comment calculer la distribution nulle dans un cas relativement simple où les observations sont supposées être i.i.d selon la loi de probabilité de Bernouilli $\mathcal{B}(y)$. Une démarche similaire peut être effectuée pour des données i.i.d. selon d'autres lois de probabilités, par exemple gaussienne ou autre. Comme les calculs peuvent être longs et compliqués on va juste établir la listes des tests les plus utilisés. Des programmes peuvent ensuite être utilisés pour obtenir directement la p-valeur correspondant aux données, ce qui nous évite de calculer la statistique correspondante et sa distribution nulle. Les tests les plus utilisés sont résumés dans le tableau suivant :

Modèle	$\mathcal{H}_0$	Nom du test
$X \sim \mathcal{B}(y)$	$y = y_0$	Test binomial
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\mu = \mu_0$	Test t (de student) à un échantillon
$X \sim \mathcal{N}(\mu_X, \sigma^2)\,, \quad Y \sim \mathcal{N}(\mu_Y, \sigma^2)$	$\mu_X = \mu_Y$	Test t (de student) à deux échantillons
$(X, Y) \sim \big(\mathcal{N}(\mu_X, \sigma^2), \mathcal{N}(\mu_Y, \sigma^2)\big)$	$\mu_X = \mu_Y$	Test t (de student) apparié (paired t-test)
$Y = a + bX + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$	$b = b_0$ ou $a = a_0$	Test de régression linéaire
$Y \in \{0, 1\}\,, \quad p(Y = 1\|X) = \frac{1}{1+e^{-(a+bX)}}$	$b = b_0$ ou $a = a_0$	Test de régression logistique

Exercice 4.3 Pour chacune des situations suivantes, indiquer quel test pourrait être utile et ce que représente les variables aléatoires, et quelle est l'hypothèse nulle.

(a) On souhaite savoir si l'exposition au soleil protège d'une certaine maladie. On mesure le temps d'exposition au soleil d'un certain nombres de patient sur une certaines période et on regarde lesquels tombent malade.

(b) Pour évaluer l'efficacité d'un régime alimentaire pour la perte de poids, on pèse un certain nombre de personnes avant et après le régime

(c) Pour évaluer si les espèces animales de petite taille ont généralement une vie plus courte, on mesure la taille et la durée de vie de diverses espèce animales.

(d) Pour savoir si un gène est impliqué dans la croissance d'une plante, on inactive ce gène dans certaines plantes et on le laisse activé dans d'autres. On mesure ensuite la taille des différentes plantes.

(e) On veur savoir si les enfant de 8 ans d'une population donnée sont en moyenne plus grand que 130 centimètres.

(f) On veut savoir si la performance au test de QI de la population d'un pays a augmenté ou diminué ces 50 dernières années.

Exercice 4.4 - La correction de Bonferroni Au cours de son travail de doctorat visant à identifier des gènes impliqué dans la communication des plantes, un chercheur effectue 50 expériences visant chacune à tester un gène différent. Avec un seuil de significativité $\alpha = 5\%$, le doctorat obtient un résultat positif sur l'ensemble de ses expériences. Très fier, il partage sa découverte avec sa directrice de thèse qui tempère aussitôt son enthousiasme. Pourquoi ? Comment le doctorant peut-il s'assurer que sa découverte est bien réelle ?

Exercice 4.5 Expliquer pourquoi il est toujours possible d'utiliser un test de régression linéaire à la place d'un test t à deux échantillons non appariés.

4.2.5 Normalité des données

Aussi bien les tests t que les tests de régression linéaire font des hypothèses de normalité des données. Cela signifie que ces tests ne doivent pas être utilisés si les données (ou les résidus dans le cas de la régression linéaire) ne sont pas distribuées selon une loi normale. Il existe des *tests de normalité* qui permettent de déterminer si un échantillon suit une loi normale, mais on peut aussi vérifier ceci graphiquement à l'aide d'un *diagramme quantile-quantile* (QQ plot, en anglais). Si notre échantillon consiste en n points, ce diagramme s'obtient en déterminant les quantiles $\frac{i}{n}$ pour $i = 1, \ldots, n$ selon une distribution gaussienne de même moyenne et variance que notre échantillon. On obtient ainsi n valeurs q_i. On réordonne les x_i par ordre croissant et on trace le graphe des points (q_i, x_i). Si ces points sont alignés sur la diagonale, la distribution des x_i est gaussienne (autrement dit normale). Si les déviation par rapport à la diagonales sont importante, les x ne sont pas distribués selon une loi normale.

4.2.6 Tests non-paramétriques

Les tests non-paramétriques sont des tests statistiques qui ne font pas d'hypothèse sur la distribution des données. En renonçant à de telles hypothèses, il est possible d'appliquer ce tests à beaucoup plus de situations, mais par contre de tels tests seront auront une *puissance* moindre que des tests paramétriques, c'est-à-dire qu'il faudra plus de données pour détecter des différences entre deux jeux de données. Autrement dit, les p-valeurs obtenues seront en général plus élevées (donc moins significatives). Elles seront toutefois plus fiables car elles font moins d'hypothèses sur les données.

Nom du test	$\mathcal{H}_0$	Usage
Mann-Whitney-Wilcoxon	$P(X > Y) \neq 0.5$	Comparaison de deux v.a.
Test exact de Fisher	$X \perp\!\!\!\perp Y$	Indépendance de deux v.a. binaires
Test du χ^2	$X \perp\!\!\!\perp Y$	Indépendance de deux v.a. catégorielles

Exemple 2.6 (suite) On souhaite savoir si les chances de survie des enfants passagers du Titanic étaient différentes entre ceux voyageant en première classe et ceux voyageant en seconde classe. On peut résumer les données selon le tableau suivant

	survivant	victime
1^e classe	5	1
2^e classe	24	0

Le *test exact de Fisher* permet de déterminer si les deux variables aléatoires sont indépendantes, autrement dit les rapport entre les deux lignes sont les mêmes pour les deux colonnes (ou vice-versa). Dans notre exemple, cela revient à demander si une victime pour 5 survivants (en première classe) et vraiment différent d'aucune victime pour 24 survivants (en deuxième classe). La p-valeur obtenue vaut 0.2 ce qu'on ne peut pas rejeter l'hypothèse nulle que les deux variables sont indépendantes.
Par contre si on fait le même test pour les femmes, on obtient le tableau suivant :

	survivante	victime
1^e classe	140	4
2^e classe	80	13

La p-valeur correspondante est alors 0.0015, ce qui est significatif en prenant un seuil de significativité à 1%. On peut donc rejeter l'hypothèse nulle et en conclure que les femmes voyageant en première classe avait plus de chance de survivre que les femmes voyageant en deuxième classe, ce qui n'est pas étonnant sachant qu'elles avaient la priorité pour l'accès au bateaux de secours.

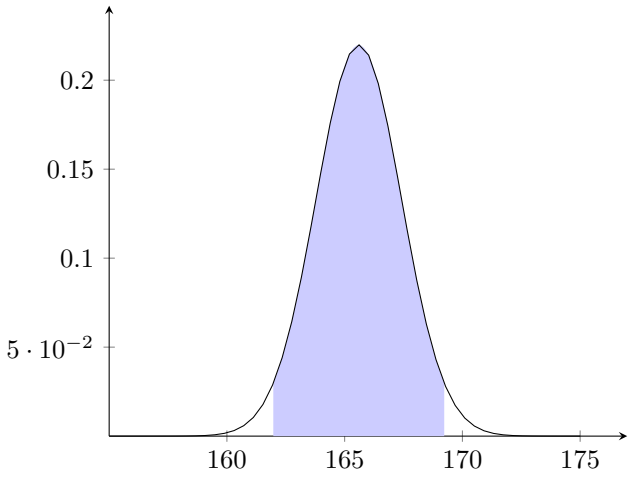
4.2.7 Types d'erreur

On peut distinguer deux types d'erreur lorsqu'on effectue un test statistique. Premièrement, on peut rejeter $\mathcal{H}_0$ alors qu'elle est vraie, c'est ce qu'on appelle un faux positif, ou une erreur de type 1. Deuxièmement, on peut ne pas rejeter $\mathcal{H}_0$ alors qu'elle est fausse, c'est ce qu'on appelle un faux négatif, ou une erreur de type 2. Ces deux types d'erreur auront dans la pratique des répercussions différentes. Par exemple, si le test correspond à un test diagnostique d'une maladie, le fait de manquer le diagnostique a des conséquences différentes que le fait de diagnostiquer à tort une maladie inexistante. Le contexte du test d'hypothèse doit donc être pris en compte pour fixer le seuil de significativité et interpréter le résultat du test. En baissant le seuil de significativité, on va réduire les faux positifs, mais augmenter les faux négatifs.

4.3 Intervalle de confiance

Les différentes manières vues au chapitre précédent pour estimer un paramètre en fonction des données (par exemple l'équation 3.13), peuvent également être considérées comme des statistiques. En effet elles sont obtenues en effectuant des opérations sur des variables aléatoires (ou sur leur réalisation). On peut donc théoriquement associer des distributions à ces estimations. Le théorème central limite présenté à la section 2.2.8 indique que si les observations sont i.i.d, alors la distribution de ces estimations peut être approximée par une gaussienne dont on peut calculer la variance théorique et donc l'écart-type. On peut ensuite utiliser cet écart-type pour construire un *intervalle de confiance* qui contiendra, avec une probabilité choisie, la vraie valeur du paramètre estimé.

Exemple 1.1 (suite) On fait l'hypothèse que les joueuses de l'équipe de foot sont représentatives, en terme de taille, de la population féminine adulte locale. La moyenne de cet échantillon vaut $\hat{\mu} = 165.6cm$, alors que l'estimation (non-biaisée) de la variance vaut $\hat{\sigma}^2 = 49.4cm^2$. Le théorème central limite nous dit qui si l'espérance de la taille de la population vaut μ et sa variance σ^2, alors en prenant 15 personnes aléatoirement dans cette population et en calculant la moyenne des tailles, on obtiendra une valeur distribuée selon une loi (presque) gaussienne dont l'espérance sera μ et dont la variance sera $\frac{\sigma^2}{15}$. En approximant σ^2 par $\hat{\sigma}^2$ et μ par $\hat{\mu}$, on obtient une loi gaussienne d'espérance 165.6 et de variance 49.4/15, telle que représentées ci-dessous.



La partie colorée correspond à l'intervalle $[\hat{\mu} - 2\frac{\hat{\sigma}}{\sqrt{15}}, \hat{\mu} + 2\frac{\hat{\sigma}}{\sqrt{15}}]$, qui a 95.5% de probabilité de contenir μ si nos hypothèses sont respectées. Autrement dit, si on applique cette procédure, on a 95% de chance que l'intervalle ainsi défini contienne μ.

L'intervalle de confiance à 95.5% pour l'estimation d'une moyenne basée sur n échantillon et donc donné par

$$[\bar{x} - 2\frac{s}{\sqrt{n}}, \bar{x} + 2\frac{s}{\sqrt{n}}], \tag{4.3}$$

où $s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ est l'estimation non biaisée de la variance des x_i. Si on souhaite un intervalle de confiance correspondant à un pourcentage différent, il suffit de remplacer le 2 par une valeur c correspondant au quantile désiré de la distribution normale (par exemple 2.58 pour un intervalle à 99%, 3 pour un intervalle à 99.73%).

Exercice 4.6 On cherche à déterminer le salaire moyen en Suisse. Pour ceci, on sélectionne aléatoirement 16 personnes salariées et auquelles on demande le salaire brut (à plein temps). On obtient les résultats suivants :

3750	6450	7830	2960	5400	25000	12200	4250	9800	6120	3450	7100
4500	5260	6300	3750								

(a) Estimer l'intervalle de confiance à 95.5% pour le salaire moyen selon ces données.

(b) Cette estimation est-elle fiable ? Expliquer pourquoi.

Exercice 4.7 On considère le modèle linéraire donnée par l'équation 2.29, est l'estimation des paramètres a et b donnée par l'équation 3.13. Proposer un intervalle de confiance à 95.5% pour ces paramètres.

4.4 Corrélations et causalité

4.5 Bayes factor

[le procès Letby]

[maybe]

4.6 Solutions des exercices

Exercice 4.1

(a) $p(T = t|\mathcal{H}_0) = \frac{1}{2^n}\binom{n}{t} = \frac{n!}{t!(n-t)!2^n}$.

(b) $p(T = t|\mathcal{H}_0) = a^t(1-a)^{n-t}\binom{n}{t} = \frac{a^t(1-a)^{n-t}n!}{t!(n-t)!}$. l'hypothèse nulle $\mathcal{H}_0 : y = a$. Ce test s'appelle le *test binomial*.

Exercice 4.2

(a) Oui, si y dénote la probabilité de Alya de gagner, on peut avoir $\mathcal{H}_0 : y = \frac{1}{2}$ et $\mathcal{H}_1 : y \neq \frac{1}{2}$.

(b) Non, car il y a plus de deux résultats possibles, donc on ne peut pas appliquer un test binomial.

(c) Non, car le taux d'échec s'observe directement, il n'y a pas besoin d'un test statistique pour le quantifier. Par contre, si on désire savoir si la probabilité d'échec est supérieure à 30%, alors un test binomial serait indiqué. Si y dénote la probabilité d'échec, on aurait $\mathcal{H}_0 : y = 0.3$ et $\mathcal{H}_1 : y > 0.3$.

(d) Oui, si y dénote la probabilité de défaut de fabrication, on peut avoir $\mathcal{H}_0 : y = 1\%$ et $\mathcal{H}_1 : y < 1\%$.

Exercice 4.3

(a) On peut utiliser un test de régression logistique avec Y décrivant le fait de tomber malade ou pas en fonction du temps d'exposition X. L'hypothèse nulle serait alors $\mathcal{H}_0 : b = 0$.

(b) On peut utiliser une test t de student apparié, c'est à dire en prenant la différence pour chaque personne entre son poids après et avant le régime et en testant $\mathcal{H}_0 : \mu = 0$, $\mathcal{H}_1 : \mu > 0$

(c) On peut utiliser un test de régression linéaire de la durée de vie en fonction de la taille.

(d) On peut effectuer un test t de student à deux échantillons non-appariés ou (un test de Mann-Whitney-Wilcoxon)

(e) On peut effectuer un test t de student à un échantillon

(f) On peut effectuer un test de régression linéaire des résultats du test en fonction des années, pour une population similaire.

Exercice 4.4 Le seuil de significativité à 5%, indique que l'hypothèse nulle sera rejetée dans 5% des cas, même si elle est vraie. Comme le chercheur à fait 50 expériences, on s'attends, si toutes les hypothèses nulles sont vraies, à ce qu'il en rejette quand même 2.5 en moyenne. C'est donc "normal" qu'il obtienne une résultat positif sur autant d'essais, et ne signifie pas qu'il a vraiment découvert quelque chose.
Pour s'assurer que sa découverte est bien réelle, le doctorant peut choisir une seuil de significativité à $\frac{5}{50}\% = 0.1\%$. Ainsi il a en tout 5% de chance d'avoir une hypothèse nulle parmi les 50 qui est rejetée si toutes les hypothèses nulles sont vraies. Le fait de diviser la seuil de significativité par le nombre d'hypothèses nulles testées s'appelle la *correction de Bonferroni*.

Exercice 4.5 Si x_i sont les n observations correspondant au premier échantillon et y_i celles correspondant au second échantillon (non apparié), on peut réunir ces deux échantillons en un seul en définissant $z_i = x_i$ et $u_i = 0$ si $i \leq n$ et et $z_i = y_{i-n}$ et $u_i = 1$ si $i > n$. On peut alors définir le modèle linéaire $u_i = a + bz_i + \epsilon_i$. L'hypothèse nulle $\mathcal{H}_0 : \mu_X = \mu_Y$ qui est testée par le test t est alors équivalente à l'hypothèse nulle $\mathcal{H}_0 : b = 0$ testée par le test de régression linéaire. L'hypothèse faite par le test t de la distribution gaussienne des échantillons se retrouve dans l'hypothèse de la distribution gaussienne des résidus faite par la régression linéaire.

Exercice 4.6

(a) L"intervalle obtenu est $[3902, 10362]$.(moyenne :7132, Variance : 28687607, écart-type de l'échantillon :5356, écart-type de la moyenne : 1615)

(b) Cet intervalle a de forte chance de contenir la vraie valeur du salaire moyen. Toutefois, l'échantillon est sans doute trop faible pour obtenir une bonne estimation de la variances des salaires. En effet, la structure des salaires est marquée par le fait qu'il y a une nombre relativement faible de très hauts salaires qui influencent la variance des salaires. Si notre échantillon contient un tel haut salaire (ici 25000.-), alors la variances sera sur-estimée car ces salaires ont une fréquence plus faible que $1/16$. Au contraire, il l'échantillon n'en contient pas, alors la variance sera sous-estimée car elle ne tient pas compte de ces hauts salaires.

Chapitre 5

Prédiction

5.1 Utiliser un modèle pour faire des prédictions

5.2 Complexité et overfitting

5.2.1 Données d'entraînement de test

5.2.2 Validation croisée

5.3 Méthodes heuristiques

5.3.1 Algorithme des k plus proches voisins

5.4 Solutions des exercices

Table des matières

Index