HEC MONTRÉAL

Modélisation statistique

Léo Belzile

Table des matières

Bienvenue

Ce cours traite de modélisation des données. Une citation célèbre attribuée à George Box dit que

> tous les modèles sont faux, mais certains sont utiles.

Ce point de vue est réducteur; McCullagh et Nelder (1989) (traduction libre) expliquent dans le préambule de leur livre

> La modélisation en science demeure, du moins partiellement, un art. Certains principes existent, en revanche, pour guider le modélisateur. Le premier est que tous les modèles sont faux; mais que **certains sont meilleurs** et **le modélisateur doit chercher le meilleur à sa portée**. En même temps, il est sage de reconnaître que la quête perpétuelle de la vérité n'est pas envisageable.

Et David R. Cox (traduction libre), de rajouter

> ...il n'est pas utile de simplement énoncer que tout modèle est faux. L'idée même de modèle sous-tend une notion de simplification et d'idéalisation. L'idée qu'un système physique, biologique ou sociologique complexe puisse être décrit de manière exacte par quelques formules est franchement absurde. La construction de **représentations idéalisées qui capturent les aspects stables les plus importants du système** est néanmoins une partie essentielle de toute analyse scientifique et les modèles statistiques ne diffèrent pas en cela d'autres types de modèles.

Pourquoi utiliser des modèles? Paul Krugman écrivait en 2010 dans son blogue

> La réponse que je donnerais est que les modèles sont un outil énormément important pour clarifier ses pensées. Vous n'avez pas à avoir une foi aveugle en votre modèle [...] pour croire qu'en mettant sur pied une description simplifiée, mais complète du fonctionnement du système [...] vous permet de gagner

une compréhension plus sophistiquée de la situation réelle. Les personnes qui n'utilisent pas de modèles finissent par se baser sur des slogans beaucoup plus simplistes que les modèles.

Contenu du cours

L'inférence statistique a pour but de tirer des conclusions formelles à partir de données. Dans le cadre de la recherche scientifique, le chercheur formule une hypothèse, collecte des données et conclut quant à la plausibilité de son hypothèse.

On distingue deux types de jeux de données: les données **expérimentales** sont typiquement collectées en milieu contrôlé suivant un protocole d'enquête et un plan d'expérience: elles servent à répondre à une question prédéterminée. L'approche expérimentale est désirable pour éviter le «jardin des embranchements» (une allégorie signifiant qu'un chercheur peut raffiner son hypothèse à la lumière des données, sans ajustement pour des variables confondantes), mais elle n'est pas toujours réalisable: par exemple, un économiste ne peut pas modifier les taux d'intérêts pour observer les impacts sur le taux d'épargne des consommateurs. Lorsque les données ont été collectées préalablement à d'autres fins, on parle de données **observationnelles**.

Par modèle, on entendra la spécification d'une loi aléatoire pour les données et une équation reliant les paramètres ou l'espérance conditionnelle d'une variable réponse Y à un ensemble de variables explicatives $\mathbf{X}$. Ce modèle peut servir à des fins de prédiction (modèle prédictif) ou pour tester des hypothèses de recherche concernant les effets de ces variables (modèle explicatif). Ces deux objectifs ne sont pas mutuellement exclusifs même si on fait parfois une distinction entre inférence et prédiction.

Un modèle prédictif permet d'obtenir des prédictions de la valeur de Y pour d'autres combinaisons de variables explicatives ou des données futures. Par exemple, on peut chercher à prédire la consommation énergétique d'une maison en fonction de la météo, du nombre d'habitants de la maison et de sa taille. La plupart des boîtes noires utilisées en apprentissage automatique tombent dans la catégorie des modèles prédictifs: ces modèles ne sont pas interprétables et ignorent parfois la structure inhérente aux données.

Par contraste, les modèles explicatifs sont souvent simples et interprétables, et les modèles de régressions sont fréquemment utilisés pour l'inférence. On se concentrera dans ce cours sur les modèles explicatifs. Par exemple, on peut chercher à déterminer

- Est-ce que les décisions intégrées (décision combinée d'achat et de quantité) sont préférables aux décisions séquentielles (décision d'acheter, puis choix de la quantité) lors de l'achat d'un produit en ligne (Duke et Amir 2023)?

- Qu'est-ce qui est le plus distrayant pour les utilisateurs de la route: parler au cellulaire, texter en conduisant, consulter sa montre intelligente (Brodeur et al. 2021)?
- Quel est l'impact de de l'inadéquation entre l'image d'un produit et sa description (Lee et Choi 2019)?
- Qu'est-ce qui explique que les prix de l'essence soient plus élevés en Gaspésie qu'ailleurs au Québec? Un rapport de surveillance des prix de l'essence en Gaspésie par la Régie de l'énergie se penche sur la question.
- Est-ce que les examens pratiques de conduite en Grande-Bretagne sont plus faciles dans les régions à faible densité de population? Une analyse du journal britannique *The Guardian* laisse penser que c'est le cas.
- Quelle est la perception environnementale d'un emballage de carton (versus de plastique) s'il englobe un contenant en plastique (Sokolova, Krishna, et Döring 2023).
- Quel est l'impact psychologique des suggestions sur le montant de dons (Moon et VanEpps 2023)?
- Est-ce que la visioconférence réduit le nombre d'interactions et d'idée créatives générées lors d'une réunion, par rapport à une rencontre en personne (Brucks et Levav 2022)?

1 Introduction

Ce chapitre couvre des rappels mathématiques de probabilité et statistique d'ordinaire couverts dans un cours de niveau collégial ou préuniversitaire.

1.1 Population et échantillons

Ce qui différencie la statistique des autres sciences est la prise en compte de l'incertitude et de la notion d'aléatoire. Règle générale, on cherche à estimer une caractéristique d'une population définie à l'aide d'un échantillon (un sous-groupe de la population) de taille restreinte.

La **population d'intérêt** est un ensemble d'individus formant la matière première d'une étude statistique. Par exemple, pour l'Enquête sur la population active (EPA) de Statistique Canada, « la population cible comprend la population canadienne civile non institutionnalisée de 15 ans et plus ». Même si on faisait un recensement et qu'on interrogeait tous les membres de la population cible, la caractéristique d'intérêt peut varier selon le moment de la collecte; une personne peut trouver un emploi, quitter le marché du travail ou encore se retrouver au chômage. Cela explique la variabilité intrinsèque.

En général, on se base sur un **échantillon** pour obtenir de l'information parce que l'acquisition de données est coûteuse. L'**inférence statistique** vise à tirer des conclusions, pour toute la population, en utilisant seulement l'information contenue dans l'échantillon et en tenant compte des sources de variabilité. Le sondeur George Gallup (traduction libre) a fait cette merveilleuse analogie entre échantillon et population:

> «Il n'est pas nécessaire de manger un bol complet de soupe pour savoir si elle est trop salé; pour autant qu'elle ait été bien brassée, une cuillère suffit.»

Un **échantillon** est un sous-groupe d'individus de la population. Si on veut que ce dernier soit représentatif, il devrait être tiré aléatoirement de la population, ce qui nécessite une certaine connaissance de cette dernière. Au siècle dernier, les bottins téléphoniques pouvaient servir à créer des plans d'enquête. C'est un sujet complexe et des cours entiers d'échantillonnage y sont consacrés. Même si on ne collectera pas de données, il convient de noter la condition essentielle pour pouvoir tirer des conclusions fiables à partir d'un

échantillon: ce dernier doit être représentatif de la population étudiée, en ce sens que sa composition doit être similaire à celle de la population, et aléatoire. On doit ainsi éviter les biais de sélection, notamment les échantillons de commodité qui consistent en une sélection d'amis et de connaissances.

Si notre échantillon est **aléatoire**, notre mesure d'une caractéristique d'intérêt le sera également et la conclusion de notre procédure de test variera d'un échantillon à l'autre. Plus la taille de ce dernier est grande, plus on obtiendra une mesure précise de la quantité d'intérêt. L'exemple suivant illustre pourquoi le choix de l'échantillon est important.

Exemple 1.1 (Gallup et l'élection présidentielle américaine de 1936)**.** Désireuse de prédire le résultat de l'élection présidentielle américaine de 1936, la revue *Literary Digest* a sondé 10 millions d'électeurs par la poste, dont 2.4 millions ont répondu au sondage en donnant une nette avance au candidat républicain Alf Landon (57%) face au président sortant Franklin D. Roosevelt (43%). Ce dernier a néanmoins remporté l'élection avec 62% des suffrages, une erreur de prédiction de 19%. Le plan d'échantillonnage avait été conçu en utilisant des bottins téléphoniques, des enregistrements d'automobiles et des listes de membres de clubs privés, etc.: la non-réponse différentielle et un échantillon biaisé vers les classes supérieures sont en grande partie responsable de cette erreur.

Gallup avait de son côté correctement prédit la victoire de Roosevelt en utilisant un échantillon aléatoire de (seulement) 50 000 électeurs. Vous pouvez lire l'histoire complète (en anglais).

1.2 Types de variables

Le résultat d'une collecte de données est un tableau, ou base de données, contenant sur chaque ligne des observations et en colonne des variables. Le Tableau 1.1 donne un exemple de structure.

- Une **variable** représente une caractéristique de la population d'intérêt, par exemple le sexe d'un individu, le prix d'un article, etc.
- une **observation**, parfois appelée donnée, est un ensemble de mesures collectées sous des conditions identiques, par exemple pour un individu ou à un instant donné.

Tableau 1.1: Premières lignes de la base de données `renfe`, qui contient les prix de 10K billets de train entre Barcelone et Madrid. Les colonnes `prix` et `duree` sont des variables numériques continues, les autres des variables catégorielles.

prix	type	classe	tarif	dest	duree	jour
143.4	AVE	Preferente	Promo	Barcelone-Madrid	190	6
181.5	AVE	Preferente	Flexible	Barcelone-Madrid	190	2
86.8	AVE	Preferente	Promo	Barcelone-Madrid	165	7
86.8	AVE	Preferente	Promo	Barcelone-Madrid	190	7
69.0	AVE-TGV	Preferente	Promo	Barcelone-Madrid	175	4

Le choix de modèle statistique ou de test dépend souvent du type de variables collectées. Les variables peuvent être de plusieurs types: quantitatives (discrètes ou continues) si elles prennent des valeurs numériques, qualitatives (binaires, nominales ou ordinales) si elles peuvent être décrites par un adjectif; je préfère le terme catégorielle, plus évocateur.

La plupart des modèles avec lesquels nous interagirons sont des modèles dits de régression, dans lesquelles on modélisation la moyenne d'une variable quantitative en fonction d'autres variables dites explicatives. Il y a deux types de variables numériques:

- une variable discrète prend un nombre dénombrable de valeurs; ce sont souvent des variables de dénombrement ou des variables dichotomiques.
- une variable continue peut prendre (en théorie) une infinité de valeurs, même si les valeurs mesurées sont arrondies ou mesurées avec une précision limitée (temps, taille, masse, vitesse, salaire). Dans bien des cas, nous pouvons considérer comme continues des variables discrètes si elles prennent un assez grand nombre de valeurs.

Les variables catégorielles représentent un ensemble fini de possibilités. On les regroupe en deux types, pour lesquels on ne fera pas de distinction:

- nominales s'il n'y a pas d'ordre entre les modalités (sexe, couleur, pays d'origine) ou
- ordinale (échelle de Likert, tranche salariale).

La codification des modalités des variables catégorielle est arbitraire; en revanche, on préservera l'ordre lorsqu'on représentera graphiquement les variables ordinales. Lors de l'estimation, chaque variable catégorielle doit est transformée en un ensemble d'indicateurs binaires 0/1: il est donc essentiel de déclarer ces dernières dans votre logiciel statistique, surtout si elles sont parfois encodées dans la base de données à l'aide de valeurs entières.

1.3 Variables aléatoires

Suppsons qu'on cherche à décrire le comportement d'un phénomène aléatoire. Pour ce faire, on cherche à décrire l'ensemble des valeurs possibles et leur probabilité/fréquence relative au sein de la population: ces dernières sont encodées dans la loi de la variable aléatoire.

On dénote les variables aléatoires par des lettres majuscules, et leurs réalisations par des minuscules: par exemple, $Y \sim \mathsf{normale}(\mu, \sigma^2)$ indique que Y suit une loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$. On parle de famille de lois si la valeur des paramètres ne sont pas spécifiées; si on fixe plutôt ces dernière, on obtient une représentation qui encode les probabilité.

Définition 1.1 (Fonctions de répartition, de masse et de densité)**.** La **fonction de répartition** $F(y)$ donne la probabilité cumulative qu'un événement n'excède pas une variable donnée, $F(y) = \mathsf{Pr}(Y \leq y)$. Si la variable Y prend des valeurs discrètes, alors on utilise la **fonction de masse** $f(y) = \mathsf{Pr}(Y = y)$ qui donne la probabilité pour chacune des valeurs de y. Si la variable Y est continue, aucune valeur numérique de y n'a de probabilité non-nulle et $\Pr(Y = y) = 0$ pour toute valeur réelle y; la **densité**, aussi dénotée $f(x)$, est une fonction est non-négative et satisfait $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$: elle décrit la probabilité d'obtenir un résultat dans un ensemble donné des réels $\mathbb{R}$, pour n'importe lequel intervalle. La densité sert à estimer la probabilité que la variable continue Y appartienne à un ensemble B, via $\mathsf{Pr}(Y \in B) = \int_B f(y)\mathrm{d}y$; la fonction de répartition est ainsi définie comme $F(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x$.

Un premier cours de statistique débute souvent par la présentation de statistiques descriptives comme la moyenne et l'écart-type. Ce sont des estimateurs des moments (centrés), qui caractérisent la loi du phénomène d'intérêt. Dans le cas de la loi normale unidimensionnelle, qui a deux paramètres, l'espérance et la variance caractérisent complètement le modèle.

Définition 1.2 (Moments)**.** Soit Y une variable aléatoire de fonction de densité (ou de masse) $f(x)$. On définit l'espérance d'une variable aléatoire Y comme

$$\mathsf{E}(Y) = \int_{\mathbb{R}} yf(y)\mathrm{d}y.$$

L'espérance est la « moyenne théorique», ou moment de premier ordre : dans le cas discret, $\mu = \mathsf{E}(Y) = \sum_{y \in \dagger} y\mathsf{Pr}(y = y)$, où $\mathcal{Y}$ représente le support de la loi, à savoir les valeurs qui peuvent prendre Y. Plus généralement, l'espérance d'une fonction $g(y)$ pour une variable

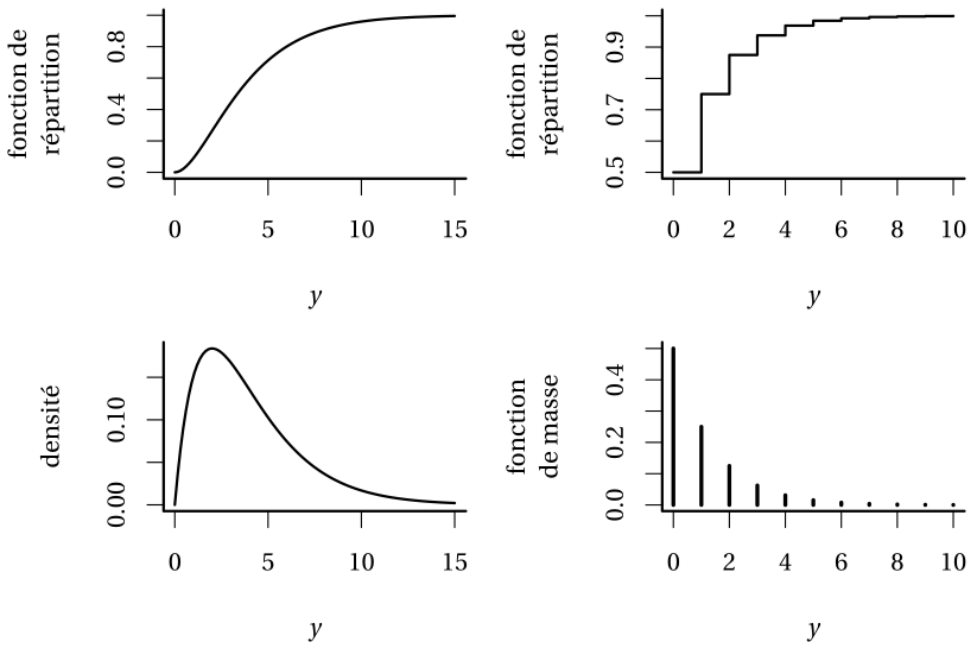

Figure 1.1: Fonctions de répartition (panneau supérieur) et fonctions de densité et de masse (panneau inférieur) pour une loi continue (gauche) et discrète (droite).

aléatoire Y est simplement l'intégrale de $g(y)$ pondérée par la densité $f(y)$. De même, si l'intégrale est convergente, la **variance** est

$$\begin{aligned}\mathsf{Va}(Y) &= \int_{\mathbb{R}} (y-\mu)^2 f(y) \mathrm{d}y \\ &= \mathsf{E}\{Y - \mathsf{E}(Y)\}^2 \\ &= \mathsf{E}(Y^2) - \{\mathsf{E}(Y)\}^2.\end{aligned}$$

L'écart-type est défini comme la racine carrée de la variance, $\mathsf{sd}(Y) = \sqrt{\mathsf{Va}(Y)}$: elle est exprimé dans les mêmes unités que celle de Y et donc plus facilement interprétable.

La notion de moments peut être généralisé à des vecteurs. Si $\boldsymbol{Y}$ est un n-vecteur, comprenant par exemple dans le cadre d'une régression des mesures d'un ensemble d'observations, alors l'espérance est calculée composante par composante,

$$\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \begin{pmatrix} \mathsf{E}(Y_1) & \cdots & \mathsf{E}(Y_n) \end{pmatrix}^\top$$

tandis que la matrice $n \times n$ de deuxième moments centrés de $\boldsymbol{Y}$, dite matrice de variance ou matrice de **covariance**, est

$$\mathsf{Va}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = \begin{pmatrix} \mathsf{Va}(Y_1) & \mathsf{Co}(Y_1, Y_2) & \cdots & \mathsf{Co}(Y_1, Y_n) \\ \mathsf{Co}(Y_2, Y_1) & \mathsf{Va}(Y_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathsf{Co}(Y_n, Y_1) & \mathsf{Co}(Y_n, Y_2) & \cdots & \mathsf{Va}(Y_n) \end{pmatrix}$$

Le ie élément diagonal de $\boldsymbol{\Sigma}$, $\sigma_{ii} = \sigma_i^2$, est la variance de Y_i, tandis que les éléments hors de la diagonale, $\sigma_{ij} = \sigma_{ji}$ $(i \neq j)$, sont les covariances des paires

$$\mathsf{Co}(Y_i, Y_j) = \int_{\mathbb{R}^2} (y_i - \mu_i)(y_j - \mu_j) f_{Y_i, Y_j}(y_i, y_j) \mathrm{d}y_i \mathrm{d}y_j.$$

Par construction, la matrice de covariance $\boldsymbol{\Sigma}$ est symmétrique. Il est d'usage de considérer la relation deux-à-deux de variables standardisées, afin de séparer la dépendance linéaire de la variabilité de chaque composante. La **corrélation linéaire** entre Y_i et Y_j est

$$\rho_{ij} = \mathsf{Cor}(Y_i, Y_j) = \frac{\mathsf{Co}(Y_i, Y_j)}{\sqrt{\mathsf{Va}(Y_i)}\sqrt{\mathsf{Va}(Y_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

La matrice de corrélation de $\boldsymbol{Y}$ est une matrice symmétrique $n \times n$ avec des uns sur la diagonale et les corrélations des pairs hors diagonale,

$$\mathsf{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \ddots & \rho_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{pmatrix}.$$

Nous modéliserons la matrice de covariance ou de corrélation des données corrélées et longitudinales par individus du même groupe (ou du même individu pour les mesures répétées) dans le Chapitre 5.

Définition 1.3 (Corrélation linéaire de Pearson)**.** Le coefficient de corrélation linéaire entre X_j et X_k, que l'on note $r_{j,k}$, cherche à mesurer la force de la relation linéaire entre deux variables, c'est-à-dire à quantifier à quel point les observations sont alignées autour d'une droite. Le coefficient de corrélation est

$$r_{j,k} = \frac{\widehat{\mathsf{Co}}(X_j, X_k)}{\{\widehat{\mathsf{Va}}(X_j)\widehat{\mathsf{Va}}(X_k)\}^{1/2}}$$

Les propriétés les plus importantes du coefficient de corrélation linéaire r sont les suivantes:

1) $-1 \leq r \leq 1$;
2) $r = 1$ (respectivement $r = -1$) si et seulement si les n observations sont exactement alignées sur une droite de pente positive (négative). C'est-à-dire, s'il existe deux constantes a et $b > 0$ ($b < 0$) telles que $y_i = a + bx_i$ pour tout $i = 1, \ldots, n$.

Règle générale,

- Le signe de la corrélation détermine l'orientation de la pente (négative ou positive)
- Plus la corrélation est près de 1 en valeur absolue, plus les points auront tendance à être alignés autour d'une droite.
- Lorsque la corrélation est presque nulle, les points n'auront pas tendance à être alignés autour d'une droite. Il est très important de noter que cela n'implique pas qu'il n'y a pas de relation entre les deux variables. Cela implique seulement qu'il n'y a pas de **relation linéaire** entre les deux variables.

La Figure 1.2 montre bien ce dernier point: ces jeux de données ont la même corrélation linéaire (quasi-nulle) et donc la même droite de régression, mais ne sont clairement pas indépendantes puisqu'elles permettent de dessiner un dinosaure ou une étoile.

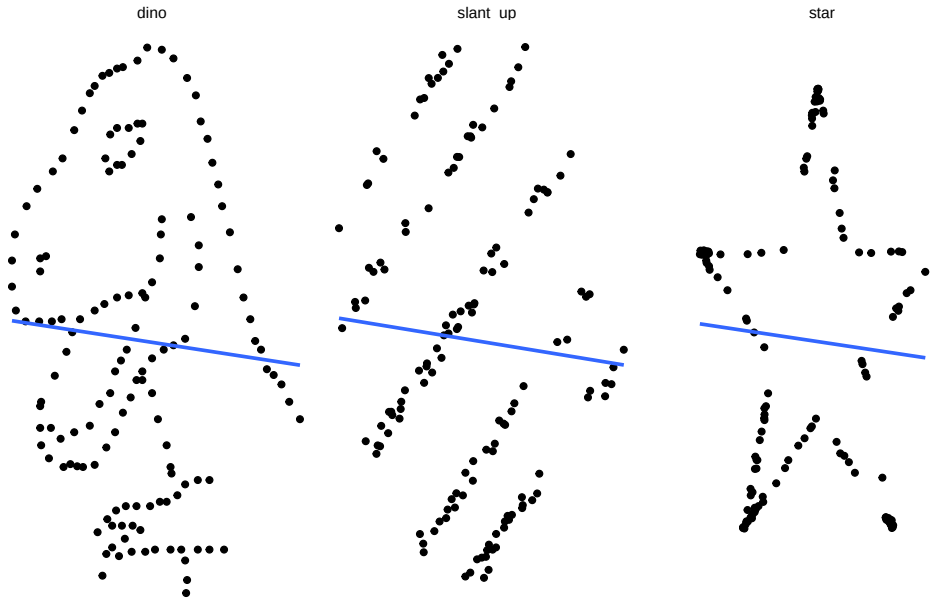

Figure 1.2: Trois jeux de données de `datasauRus`, avec une corrélation linéaire de -0.06 et des statistiques descriptives moyenne, écart-type, etc. identiques pour chaque jeu de données.

Définition 1.4 (Biais)**.** Le biais d'un estimateur $\hat{\theta}$ pour un paramètre θ est

$$\mathsf{biais}(\hat{\theta}) = \mathsf{E}(\hat{\theta}) - \theta$$

L'estimateur est non biaisé si $\mathsf{biais}(\hat{\theta}) = 0$.

Exemple 1.2 (Estimateurs sans biais)**.** L'estimateur sans biais de l'espérance de Y pour un échantillon aléatoire simple $Y_1, \ldots, Y_n$ est la moyenne empirique $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ et celui de la variance $S_n = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$.

Un estimateur sans biais est souhaitable, mais pas toujours optimal. Quelquefois, il n'existe pas d'estimateur non-biaisé pour un paramètre! Dans plusieurs cas, on cherche un estimateur qui minimise l'erreur quadratique moyenne.

Souvent, on cherche à balancer le biais et la variance: rappelez-vous qu'un estimateur est une variable aléatoire (étant une fonction de variables aléatoires) et qu'il est lui-même variable: même s'il est sans biais, la valeur numérique obtenue fluctuera d'un échantillon à l'autre.

Définition 1.5 (Erreur quadratique moyenne)**.** On peut chercher un estimateur qui minimise l'erreur quadratique moyenne,

$$\mathsf{EQM}(\hat{\theta}) = \mathsf{E}\{(\hat{\theta} - \theta)^2\} = \mathsf{Va}(\hat{\theta}) + \{\mathsf{E}(\hat{\theta})\}^2.$$

Cette fonction objective est donc un compromis entre le carré du biais et la variance de l'estimateur.

La plupart des estimateurs que nous considérerons dans le cadre du cours sont des estimateurs du maximum de vraisemblance. Ces derniers sont asymptotiquement efficaces, c'est-à-dire qu'ils minimisent l'erreur quadratique moyenne parmi tous les estimateurs possibles quand la taille de l'échantillon est suffisamment grande. Ils ont également d'autre propriétés qui les rendent attractifs comme choix par défaut pour l'estimation. Il ne sont pas nécessairement sans biais

1.4 Loi discrètes

Plusieurs lois aléatoires décrivent des phénomènes physiques simples et ont donc une justification empirique; on revisite les distributions ou loi discrètes les plus fréquemment couvertes.

Définition 1.6 (Loi de Bernoulli)**.** On considère un phénomène binaire, comme le lancer d'une pièce de monnaie (pile/face). De manière générale, on associe les deux possibilités à succès/échec et on suppose que la probabilité de "succès" est p. Par convention, on représente les échecs (non) par des zéros et les réussites (oui) par des uns. Donc, si la variable Y vaut 0 ou 1, alors $\mathsf{Pr}(Y = 1) = p$ et la probabilité complémentaire est $\mathsf{Pr}(Y = 0) = 1 - p$. La fonction de masse de la loi Bernoulli s'écrit de façon plus compacte

$$\mathsf{Pr}(Y = y) = p^y(1 - p)^{1-y}, \quad y = 0, 1.$$

Un calcul rapide montre que $\mathsf{E}(Y) = p$ et $\mathsf{Va}(Y) = p(1 - p)$. Effectivement,

$$\mathsf{E}(Y) = \mathsf{E}(Y^2) = p \cdot 1 + (1 - p) \cdot 0 = p.$$

Voici quelques exemples de questions de recherches comprenant une variable réponse binaire:

- est-ce qu'un client potentiel a répondu favorablement à une offre promotionnelle?
- est-ce qu'un client est satisfait du service après-vente?
- est-ce qu'une firme va faire faillite au cours des trois prochaines années?
- est-ce qu'un participant à une étude réussit une tâche assignée?

Plus généralement, on aura accès à des données aggrégées.

Exemple 1.3 (Loi binomiale)**.** Si les données représentent la somme d'événements Bernoulli indépendants, la loi du nombre de réussites Y pour un nombre d'essais donné m est dite binomiale, dénotée $\mathsf{Bin}(m, p)$; sa fonction de masse est

$$\mathsf{Pr}(Y = y) = \binom{m}{y} p^y(1 - p)^{1-y}, \quad y = 0, 1.$$

La vraisemblance pour un échantillon de la loi binomiale est (à constante de normalisation près qui ne dépend pas de p) la même que pour un échantillon aléatoire de m variables Bernoulli indépendantes. L'espérance d'une variable binomiale est $\mathsf{E}(Y) = mp$ et la variance $\mathsf{Va}(Y) = mp(1 - p)$.

On peut ainsi considérer le nombre de personnes qui ont obtenu leur permis de conduire parmi m candidat(e)s ou le nombre de clients sur m qui ont passé une commande de plus de 10$ dans un magasin.

Plus généralement, on peut considérer des variables de dénombrement qui prennent des valeurs entières. Parmi les exemples de questions de recherches comprenant une variable réponse de dénombrement:

- le nombre de réclamations faites par un client d'une compagnie d'assurance au cours d'une année.
- le nombre d'achats effectués par un client depuis un mois.
- le nombre de tâches réussies par un participant lors d'une étude.

Exemple 1.4 (Loi de Poisson)**.** Si la probabilité d'un événement Bernoulli est petite et qu'il est rare d'obtenir un succès dans le sens où $mp \to \lambda$ quand le nombre d'essais m augmente, alors le nombre de succès suit approximateivement une loi de Poisson de fonction de masse

$$\mathsf{Pr}(Y = y) = \frac{\exp(-\lambda)\lambda^y}{\Gamma(y+1)}, \quad y = 0, 1, 2, \ldots$$

où $\Gamma(\cdot)$ dénote la fonction gamma, et $\Gamma(y+1) = y!$ si y est un entier. Le paramètre λ de la loi de Poisson représente à la fois l'espérance et la variance de la variable, c'est-à-dire que $\mathsf{E}(Y) = \mathsf{Va}(Y) = \lambda$.

Exemple 1.5 (Loi binomiale négative)**.** On considère une série d'essais Bernoulli de probabilité de succès p jusqu'à l'obtention de m succès. Soit Y, le nombre d'échecs: puisque la dernière réalisation doit forcément être un succès, mais que l'ordre des succès/échecs précédents n'importe pas, la fonction de masse de la loi binomiale négative est

$$\mathsf{Pr}(Y = y) = \binom{m-1+y}{y} p^m (1-p)^y.$$

La loi binomiale négative apparaît également si on considère la loi non-conditionnelle du modèle hiérarchique gamma-Poisson, dans lequel on suppose que le paramètre de la moyenne de la loi Poisson est aussi aléatoire, c'est-à-dire $Y \mid \Lambda = \lambda \sim \mathsf{Po}(\lambda)$ et Λ suit une loi gamma de paramètre de forme r et de paramètre d'échelle θ, dont la densité est

$$f(x) = \theta^{-r} x^{r-1} \exp(-x/\theta)/\Gamma(r).$$

Le nombre d'événements suit alors une loi binomiale négative.

La paramétrisation la plus courante pour la modélisation est légèrement différente: pour un paramètre $r > 0$ (pas forcément entier), on écrit la fonction de masse

$$\mathsf{Pr}(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y,$$

où Γ dénote la fonction gamma. Dans cette paramétrisation, la moyenne théorique et la variance sont $\mathsf{E}(Y) = \mu$ et $\mathsf{Va}(Y) = \mu + k\mu^2$, où $k = 1/r$. La variance d'une variable binomiale négative est *supérieure* à sa moyenne et le modèle est utilisé comme alternative à la loi de Poisson pour modéliser la surdispersion.

1.5 Lois continues

On considère plusieurs lois de variables aléatoires continues; certaines servent de lois pour des tests d'hypothèse et découlent du théorème central limite (notamment les lois normales, Student, Fisher ou F, et khi-deux).

Définition 1.7 (Loi beta)**.** La loi beta $\mathsf{Beta}(\alpha, \beta)$ est une loi sur l'intervalle $[0, 1]$ avec paramètres de forme $\alpha > 0$ et $\beta > 0$. Sa densité est

$$f(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{1-\beta}, \qquad x \in [0, 1].$$

Le cas $\alpha = \beta = 1$, dénotée également $\mathsf{unif}(0, 1)$, correspond à la loi standard uniforme.

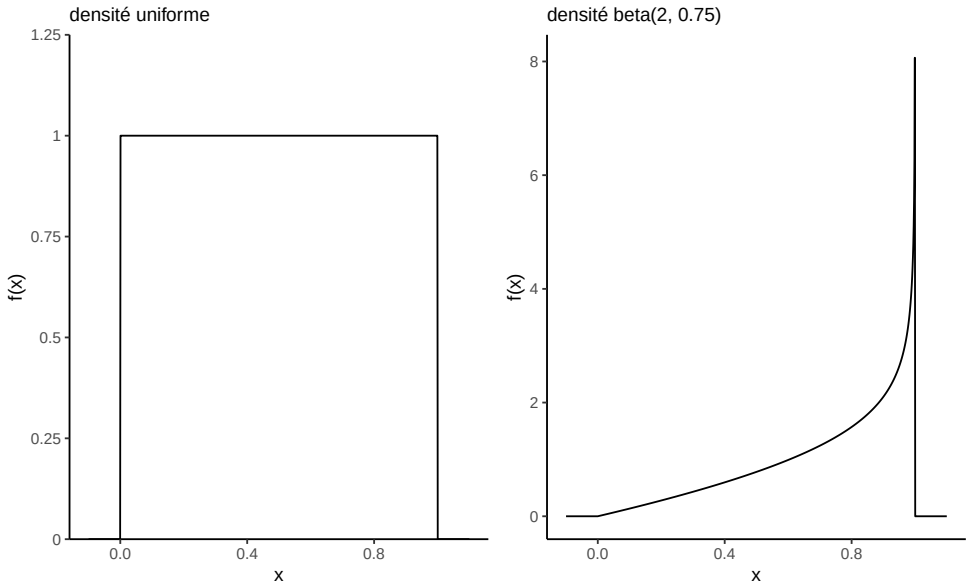

Figure 1.3: Fonctions de densité de lois uniformes et beta(2, 3/4) sur l'intervalle [0,1].

Définition 1.8 (Loi exponentielle)**.** La loi exponentielle figure de manière proéminente dans l'étude des temps d'attente pour les phénomènes Poisson et en analyse de survie. Une caractéristique clé de la loi est son absence de mémoire: $\Pr(Y \geq y + u \mid Y > u) = \Pr(Y > u)$ pour $Y > 0$ et $y, u > 0$.

La fonction de répartition de la loi exponentielle $Y \sim \mathsf{Exp}(\beta)$ où $\beta > 0$, est $F(x) = 1 - \exp(-\beta x)$ et sa fonction de densité est $f(x) = \beta \exp(-\beta x)$ pour $x > 0$. La moyenne théorique de la loi est β.

Définition 1.9 (Loi normale)**.** De loin la plus continue des distributions, la loi normale intervient dans le théorème central limite, qui dicte le comportement aléatoire de la moyenne de grand échantillons. La loi normale est pleinement caractérisée par son espérance $\mu \in \mathbb{R}$ et son écart-type $\sigma > 0$. Loi symmétrique autour de μ, c'est une famille de localisation et d'échelle. Sa fonction de densité,

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}.$$

en forme de cloche, est symmétrique autour de μ, qui est aussi le mode de la distribution.

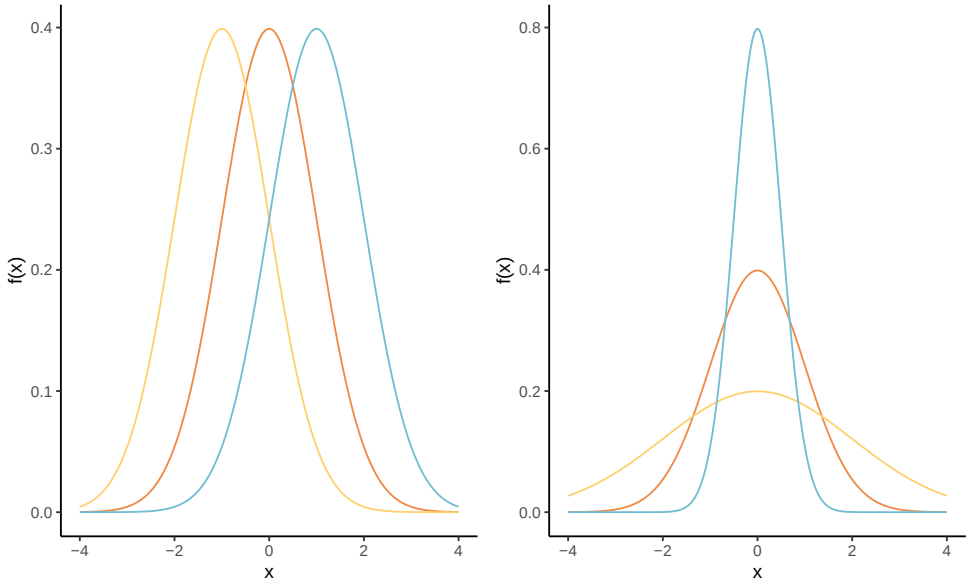

Figure 1.4: Densités de loi normales avec des paramètres de moyenne différents (gauche) et des paramètres d'échelle différents (droite).

The distribution function of the normal distribution is not available in closed-form. La loi normale est une famille de localisation échelle: si $Y \sim \mathsf{normale}(\mu, \sigma^2)$, alors $Z = (Y-\mu)/\sigma \sim \mathsf{normale}(0,1)$. Inversement, si $Z \sim \mathsf{normale}(0,1)$, alors $Y = \mu + \sigma Z \sim \mathsf{normale}(\mu, \sigma^2)$.

Nous verrons aussi l'extension multidimensionnelle de la loi normale: un d vecteur $\boldsymbol{Y} \sim \mathsf{normal}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ admet une fonction de densité égale à

$$f(\boldsymbol{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

Le vecteur de moyenne $\boldsymbol{\mu}$ contient l'espérance de chaque composante, tandis que $\boldsymbol{\Sigma}$ est la matrice de covariance de $\boldsymbol{Y}$. Une propriété unique à la loi normale (muldimensionnelle)

est le lien entre indépendance et matrice de covariance: si Y_i et Y_j sont indépendants, alors l'entrée (i,j) hors diagonale de $\boldsymbol{\Sigma}$ est nulle.

Les trois lois suivantes ne sont pas couvertes dans les cours d'introduction, mais elles interviennent régulièrement dans les cours de mathématique statistique et serviront d'étalon de mesure pour déterminer si les statistiques de test sont extrêmes sous l'hypothèse nulle.

Définition 1.10 (Loi khi-deux)**.** La loi de khi-deux avec $\nu > 0$ degrés de liberté, dénotée χ^2_ν ou $\mathsf{khi}-\mathsf{deux}(\nu)$ joue un rôle important en statistique. Sa densité est

$$f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1}\exp(-x/2), \qquad x > 0.$$

Elle est obtenue pour ν entier en prenant la somme de variables normales centrées et réduites au carré: si $Y_i \overset{\text{iid}}{\sim} \mathsf{normale}(0,1)$ pour $i = 1,\ldots,k$, alors $\sum_{i=1}^k Y_i^2 \sim \chi^2_k$. L'espérance de la loi χ^2_k est k.

Si on considère un échantillon aléatoire et identiquement distribution de n observations de lois normales, alors la variance empirique repondérée satisfait $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Définition 1.11 (Loi Student-t)**.** La loi Student-t avec $\nu > 0$ degrés de liberté est une famille de localisation et d'échelle de densité symmétrique. On la dénote $\mathsf{Student}(\nu)$ dans le cas centré réduit.

Son nom provient d'un article de William Gosset sous le pseudonyme Student (Gosset 1908), qui a introduit la loi comme approximation au comportement de la statistique t. La densité d'une loi Student standard avec ν degrés de liberté est

$$f(y;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

La loi a des ailes à décroissance polynomiale, est symmétrique autour de zéro et unimodale. Quand $\nu \to \infty$, on recouvre une loi normale, mais les ailes sont plus lourdes que la loi normale. Effectivement, seuls les $\nu-1$ premiers moments de la distribution existent: la loi $\mathsf{Student}(2)$ n'a pas de variance.

Si les n observations indépendantes et identiquement distribuées $Y_i \sim \mathsf{normale}(\mu,\sigma^2)$, alors la moyenne empirique centrée, divisée par la variance empirique, $(\overline{Y}-\mu)/S^2$, suit une loi Student-t avec $n-1$ degrés de liberté.

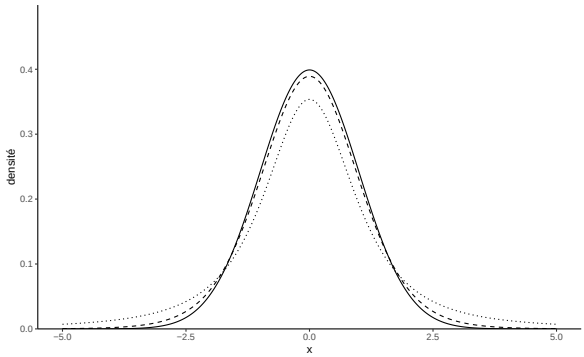

Figure 1.5: Comparaison de la densité Student-t versus normale pour différents degrés de liberté avec $\nu = 2$ (pointillé), $\nu = 10$ (traitillé) et la loi normale ($\nu = \infty$).

Définition 1.12 (Loi de Fisher)**.** La loi de Fisher, ou loi F, sert à déterminer le comportement en grand échantillon de statistiques de test pour la comparaison de plusieurs moyennes (analyse de variance) sous un postulat de normalité des observations.

La loi F, dite de Fisher et dénotée $\mathsf{Fisher}(\nu_1, \nu_2)$, est obtenue en divisant deux variables khi-deux indépendantes de degrés de liberté ν_1 et ν_2. Spécifiquement, si $Y_1 \sim \chi^2_{\nu_1}$ et $Y_2 \sim \chi^2_{\nu_2}$, alors

$$F = \frac{Y_1/\nu_1}{Y_2/\nu_2} \sim \mathsf{Fisher}(\nu_1, \nu_2)$$

La loi de Fisher tend vers une loi $\chi^2_{\nu_1}$ quand $\nu_2 \to \infty$.

1.6 Graphiques

Cette section sert à réviser les principales représentations graphiques de jeux de données selon la catégorie des variables.

Le principal type de graphique pour représenter la distribution d'une variable catégorielle est le diagramme en bâtons, dans lequel la fréquence de chaque catégorie est présentée sur l'axe des ordonnées (y) en fonction de la modalité, sur l'axe des abscisses (x), et ordonnées pour des variables ordinales. Cette représentation est en tout point supérieur au diagramme en camembert, une engeance répandu qui devrait être honnie (notamment parce que l'humain juge mal les différences d'aires, qu'une simple rotation change la perception du graphique et qu'il est difficile de mesurer les proportions) — ce n'est pas de la tarte!

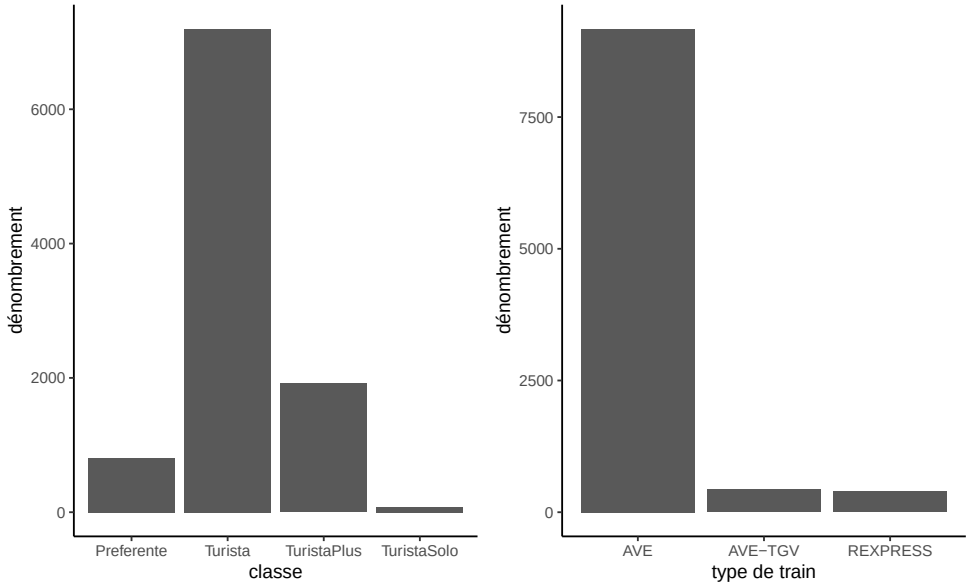

Figure 1.6: Diagramme en bâtons pour la classe des billets de trains du jeu de données Renfe.

Puisque les variables continues peuvent prendre autant de valeurs distinctes qu'il y a d'observations, on ne peut simplement compter le nombre d'occurrence par valeur unique. On regroupera plutôt dans un certain nombre d'intervalle, en discrétisant l'ensemble des valeurs en classes pour obtenir un histogramme. Le nombre de classes dépendra du nombre d'observations si on veut que l'estimation ne soit pas impactée par le faible nombre d'observations par classe: règle générale, le nombre de classes ne devrait pas dépasser $\sqrt{n}$, où n est le nombre d'observations de l'échantillon. On obtiendra la fréquence de chaque classe, mais si on normalise l'histogramme (de façon à ce que l'aire sous les bandes verticales égale un), on obtient une approximation discrète de la fonction de densité. Faire varier le nombre de classes permet parfois de faire apparaître des caractéristiques de la variable (notamment la multimodalité, l'asymmétrie et les arrondis).

Puisque qu'on groupe les observations en classe pour tracer l'histogramme, il est difficile de voir l'étendue des valeurs que prenne la variable: on peut rajouter des traits sous l'histogramme pour représenter les valeurs uniques prises par la variable, tandis que la hauteur de l'histogramme nous renseigne sur leur fréquence relative.

Définition 1.13 (Boîte à moustaches)**.** Elle représente graphiquement cinq statistiques descriptives.

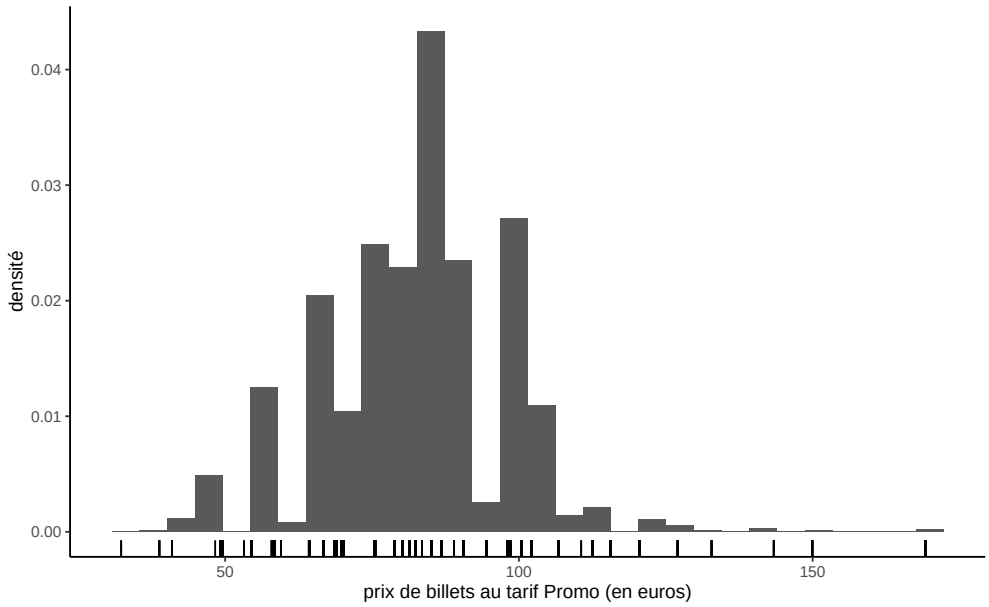

Figure 1.7: Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe

- La boîte donne les 1e, 2e et 3e quartiles q_1, q_2, q_3. Il y a donc 50% des observations sont au-dessus/en-dessous de la médiane q_2 qui sépare en deux la boîte.
- La longueur des moustaches est moins de 1.5 fois l'écart interquartile $q_3 - q_1$ (tracée entre 3e quartile et le dernier point plus petit que $q_3 + 1.5(q_3 - q_1)$, etc.)
- Les observations au-delà des moustaches sont encerclées. Notez que plus le nombre d'observations est élevé, plus le nombres de valeurs aberrantes augmente. C'est un défaut de la boîte à moustache, qui a été conçue pour des jeux de données qui passeraient pour petits selon les standards actuels.

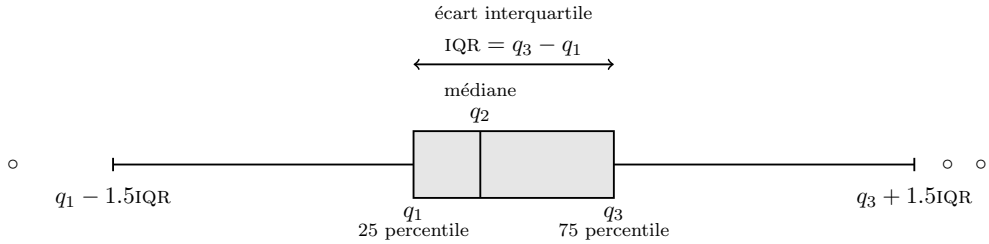

Figure 1.8: Boîte à moustache.

On peut représenter la distribution d'une variable réponse continue en fonction d'une

variable catégorielle en traçant une boîte à moustaches pour chaque catégorie et en les disposant côte-à-côte. Une troisième variable catégorielle peut être ajoutée par le biais de couleurs, comme dans la Figure 1.9.

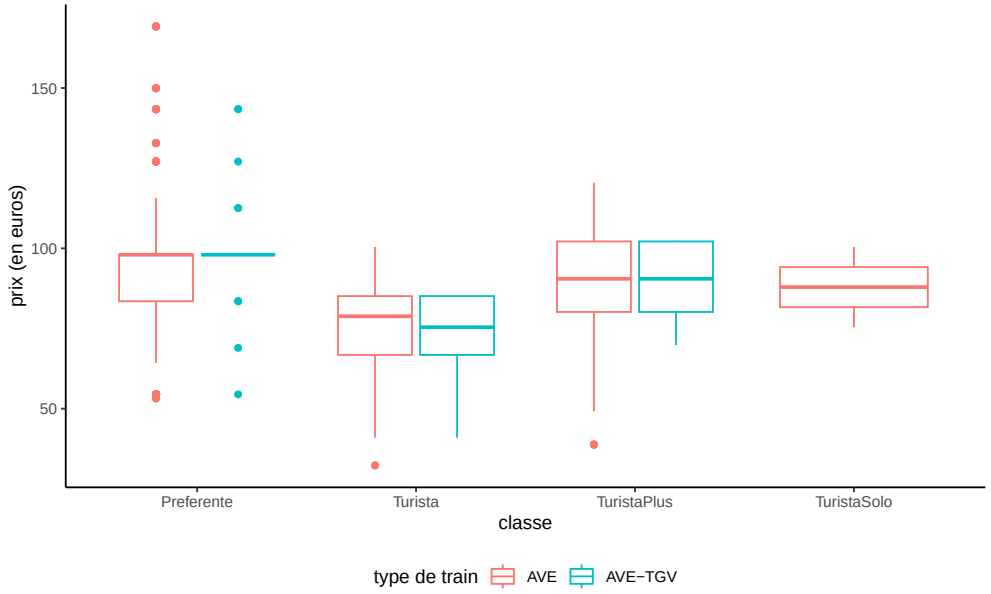

Figure 1.9: Boîte à moustaches du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe.

Si on veut représenter la covariabilité de deux variables continues, on utilise un nuage de points où chaque variable est représentée sur un axe et chaque observation donne la coordonnée des points. Si la représentation graphique est dominée par quelques valeurs très grandes, une transformation des données peut être utile: vous verrez souvent des données positives à l'échelle logarithmique. Si le nombre d'observations est très grand, il devient difficile de distinguer quoi que ce soit. On peut alors ajouter de la transparence ou regrouper des données en compartiments bidimensionnels (un histogramme bidimensionnel), dont la couleur représente la fréquence de chaque compartiment. Le paneau gauche de Figure 1.10 montre un nuage de points de 100 observations simulées, tandis que celui de droite représente des compartiments hexagonaux contenant 10 000 points.

Si on ajuste un modèle à des données, il convient de vérifier la qualité de l'ajustement et l'adéquation du modèle, par exemple graphiquement.

Définition 1.14 (Diagrammes quantiles-quantiles)**.** Le diagramme quantile-quantile sert à vérifier l'adéquation du modèle et découle du constat suivant: si Y est une variable aléatoire continue et F sa fonction de répartition, alors l'application $F(Y) \sim \mathsf{unif}(0,1)$, une loi

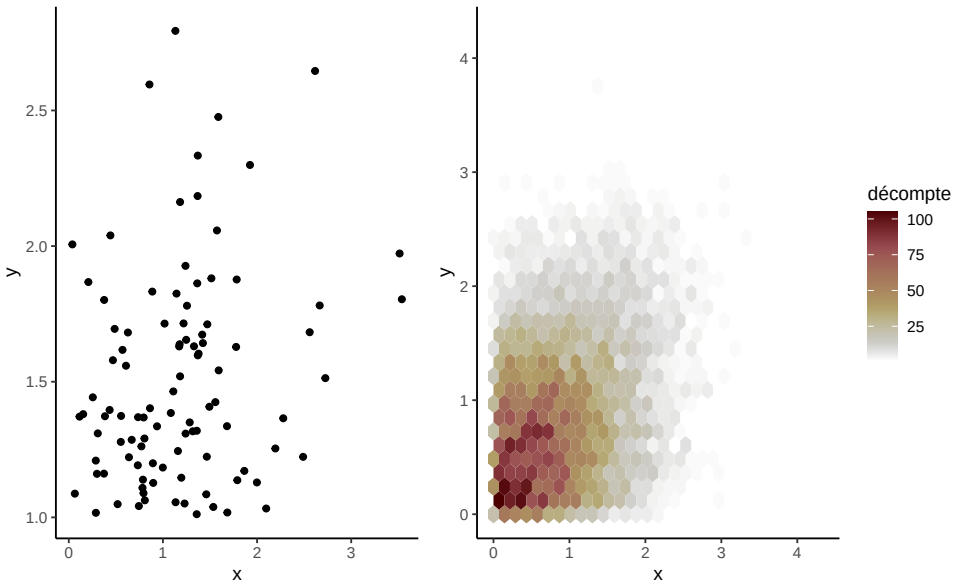

Figure 1.10: Nuage de points (gauche) et diagramme hexagonal (droite) pour des données simulées.

uniforme standard. De la même façon, appliquer la fonction quantile à une variable uniforme permet de simuler de la loi F, et donc $F^{-1}(U)$. Supposons un échantillon uniforme de taille n. On peut démontrer que, pour des variables continues, les statistiques d'ordre $U_{(1)} \leq \cdots \leq U_{(n)}$ ont une loi marginale beta, avec $U_{(k)} \sim \mathsf{Beta}(k, n+1-k)$ d'espérance $k/(n+1)$.

Les paramètres de la loi F sont inconnus, mais on peut obtenir un estimateur $\widehat{F}$ et appliquer la transformation inverse pour obtenir une variable approximativement uniforme. Un diagramme quantile-quantile représente les données en fonction des moments des statistiques d'ordre transformées

- sur l'axe des abscisses, les quantiles théoriques $\widehat{F}^{-1}\{\mathrm{rang}(Y_i)/(n+1)\}$
- sur l'axe des ordonnées, les quantiles empiriques Y_i

Si le modèle est adéquat, les valeurs ordonnées devraient suivre une droite de pente unitaire qui passe par l'origine. Le diagramme probabilité-probabilité représente plutôt les données à l'échelle uniforme $\{\mathrm{rang}(Y_i)/(n+1), \widehat{F}(Y_i)\}$.

Même si on connaissait exactement la loi aléatoire des données, la variabilité intrinsèque à l'échantillon fait en sorte que des déviations qui semblent significatives et anormales à l'oeil de l'analyste sont en fait compatibles avec le modèle: un simple estimé ponctuel

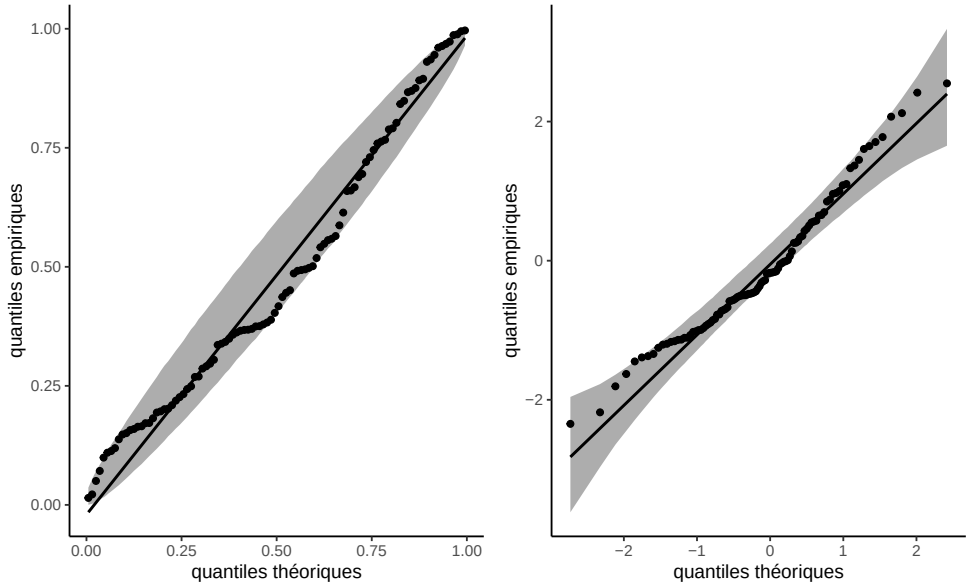

Figure 1.11: Diagramme probabilité-probabilité (gauche) et quantile-quantile normal (droite)

sans mesure d'incertitude ne permet donc pas facilement de voir ce qui est plausible ou pas. On va donc idéalement ajouter un intervalle de confiance (approximatif) ponctuel ou conjoint au diagramme.

Pour obtenir l'intervalle de confiance approximatif, la méthode la plus simple est par simulation, en répétant B fois les étapes suivantes

1. simuler un échantillon $\{Y_i^{(b)}\}(i = 1, \ldots, n)$ du modèle $\widehat{F}$
2. estimer les paramètres du modèle F pour obtenir $\widehat{F}_{(b)}$
3. calculer et stocker les positions $\widehat{F}_{(b)}^{-1}\{i/(n+1)\}$.

Le résultat de cette opération sera une matrice $n \times B$ de données simulées; on obtient un intervalle de confiance symmétrique en conservant le quantile $\alpha/2$ et $1-\alpha/2$ de chaque ligne. Le nombre de simulation B devrait être large (typiquement 999 ou davantage) et être choisi de manière à ce que B/α soit un entier.

Pour l'intervalle de confiance ponctuel, chaque valeur représente une statistique et donc individuellement, la probabilité qu'une statistique d'ordre sorte de l'intervalle de confiance est α. En revanche, les statistiques d'ordres ne sont pas indépendantes et sont qui est plus ordonnées, ce qui fait qu'un point hors de l'intervalle risque de n'être pas isolé. Les intervalles présentés dans la Figure 1.11 sont donc ponctuels. La variabilité des

statistiques d'ordre uniformes est plus grande autour de 1/2, mais celles des variables transformées dépend de F.

L'interprétation d'un diagramme quantile-quantile nécessite une bonne dose de pratique et de l'expérience: cette publication par *Glen_b* sur StackOverflow résume bien ce qu'on peut détecter ou pas en lisant le diagramme.

1.7 Loi des grands nombres

Un estimateur est dit **convergent** si la valeur obtenue à mesure que la taille de l'échantillon augmente s'approche de la vraie valeur que l'on cherche à estimer. Mathématiquement parlant, un estimateur est dit convergent s'il converge en probabilité, ou $\hat{\theta} \xrightarrow{\Pr} \theta$: en langage commun, la probabilité que la différence entre $\hat{\theta}$ et θ diffèrent est négligeable quand n est grand.

La condition *a minima* pour le choix d'un estimateur est donc la convergence: plus on récolte d'information, plus notre estimateur devrait s'approcher de la valeur qu'on tente d'estimer.

La loi des grands nombres établit que la moyenne empirique de n observations indépendantes de même espérance, $\overline{Y}_n$, tend vers l'espérance commune des variables μ, où $\overline{Y}_n \to \mu$. En gros, ce résultat nous dit que l'on réussit à approximer de mieux en mieux la quantité d'intérêt quand la taille de l'échantillon (et donc la quantité d'information disponible sur le paramètre) augmente. La loi des grands nombres est très utile dans les expériences Monte Carlo: on peut ainsi approximer par simulation la moyenne d'une fonction $g(x)$ de variables aléatoires en simulant de façon répétée des variables Y indépendantes et identiquement distribuées et en prenant la moyenne empirique $n^{-1}\sum_{i=1}^{n} g(Y_i)$.

Si la loi des grands nombres nous renseigne sur le comportement limite ponctuel, il ne nous donne aucune information sur la variabilité de notre estimé de la moyenne et la vitesse à laquelle on s'approche de la vraie valeur du paramètre.

1.8 Théorème central limite

Le théorème central limite dit que, pour un échantillon aléatoire de taille n dont les observations sont indépendantes et tirées d'une loi quelconque d'espérance μ et de variance finie σ^2, alors la moyenne empirique tend non seulement vers μ, mais à une vitesse précise:

- l'estimateur $\overline{Y}$ sera centré autour de μ,

- l'erreur-type sera de $\sigma/\sqrt{n}$; le taux de convergence est donc de $\sqrt{n}$. Ainsi, pour un échantillon de taille 100, l'erreur-type de la moyenne empirique sera 10 fois moindre que l'écart-type de la variable aléatoire sous-jacente.
- la loi approximative de la moyenne $\overline{Y}$ sera normale.

Mathématiquement, le théorème central limite dicte que $\sqrt{n}(\overline{Y}-\mu) \xrightarrow{\text{d}} \text{normale}(0,\sigma^2)$. Si n est grand (typiquement supérieur à 30, mais cette règle dépend de la loi sous-jacente de Y), alors $\overline{Y} \dot{\sim} \text{normale}(\mu,\sigma^2/n)$.

Comment interpréter ce résultat? On considère comme exemple le temps de trajet moyen de trains à haute vitesse AVE entre Madrid et Barcelone opérés par la Renfe.

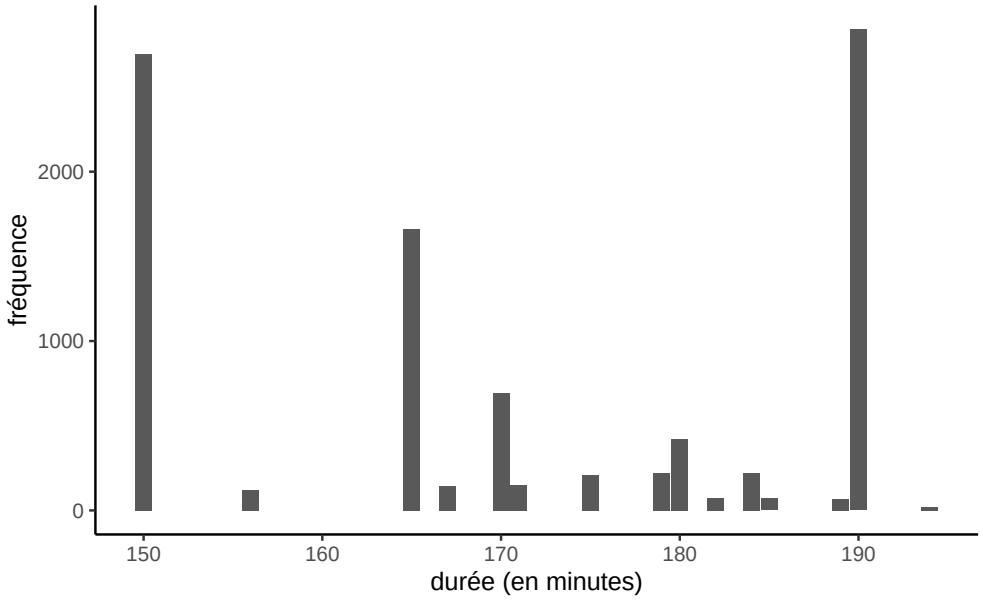

Figure 1.12: Distribution empirique des temps de trajet en trains à grande vitesse.

Une analyse exploratoire indique que la durée du trajet de la base de données est celle affichée sur le billet (et non le temps réel du parcours). Ainsi, il n'y a ainsi que 15 valeurs possibles. Le temps affiché moyen pour le parcours, estimé sur la base de 9603 observations, est de 170 minutes et 41 secondes. La Figure 1.12 montre la distribution empirique des données.

Considérons maintenant des échantillons de taille $n = 10$. Dans notre premier échantillon aléatoire, la durée moyenne affichée est 169.3 minutes, elle est de 167 minutes dans le deuxième, de 157.9 dans le troisième, et ainsi de suite.

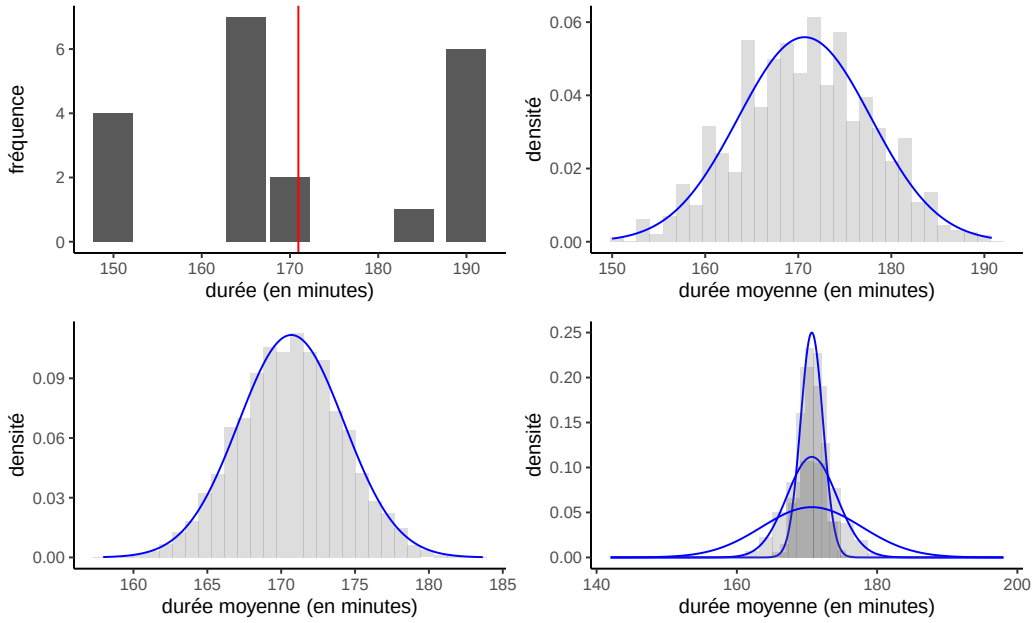

Figure 1.13: Représentation graphique du théorème central limite: échantillon aléatoire de 20 observations avec leur moyenne empirique (trait vertical rouge) (en haut à gauche). Les trois autres panneaux montrent les histogrammes des moyennes empiriques d'échantillons répétés de taille 5 (en haut à droite), 20 (en bas à gauche) et les histogrammes pour $n = 5, 20, 100$ (en bas à droite) avec courbe de densité de l'approximation normale fournie par le théorème central limite.

Supposons qu'on tire $B = 1000$ échantillons différents, chacun de taille $n = 5$, de notre ensemble, et qu'on calcule la moyenne de chacun d'entre eux. Le graphique supérieur droit de la Figure 1.13 montre un de ces 1000 échantillons aléatoire de taille $n = 20$ tiré de notre base de données. Les autres graphiques de la Figure 1.13 illustrent l'effet de l'augmentation de la taille de l'échantillon: si l'approximation normale est approximative avec $n = 5$, la distribution des moyennes est virtuellement identique à partir de $n = 20$. Plus la moyenne est calculée à partir d'un grand échantillon (c'est-à-dire, plus n augmente), plus la qualité de l'approximation normale est meilleure et plus la courbe se concentre autour de la vraie moyenne; malgré le fait que nos données sont discrètes, la distribution des moyennes est approximativement normale.

On a considéré une seule loi aléatoire inspirée de l'exemple, mais vous pouvez vous amuser à regarder l'effet de la distribution sous-jacent et de la taille de l'échantillon nécessaire pour que l'effet du théorème central limite prenne effet: il suffit pour cela de simulant des observations d'une loi quelconque de variance finie, en utilisant par exemple cette

applette.

Les statistiques de test qui découlent d'une moyenne centrée-réduite (ou d'une quantité équivalente pour laquelle un théorème central limite s'applique) ont souvent une loi nulle standard normale, du moins asymptotiquement (quand n est grand, typiquement $n > 30$ est suffisant). C'est ce qui garantie la validité de notre inférence!

2 Inférence statistique

Dans la plupart des domaines scientifiques, les donnéese empiriques issues d'expériences contribuent à l'édification de la science. Afin de tirer des conclusions en faveur ou à l'encontre d'une théorie, les chercheurs se tournent (souvent à contrecoeur) vers la statistique. Cela a conduit à la prédominance de l'utilisation du cadre des tests statistiques et à la prépondérance des valeurs-p dans les articles scientifiques, souvent employées de manière abusive ou fautive dans les articles de journaux. La falsification d'une hypothèse nulle n'est pas suffisante pour fournir des résultats substantiels pour une théorie.

Comme les cours d'introduction aux statistiques présentent généralement des tests d'hypothèses sans accorder beaucoup d'attention aux principes de construction sous-jacents de ces procédures, les utilisateurs ont souvent une vision réductrice des statistiques. Plusieurs voient les statistiques comme un catalogue de procédures pré-établies. Pour faire une analogie culinaire, les utilisateurs se concentrent sur l'apprentissage en vase clos des recettes plutôt que d'essayer de comprendre les bases de la cuisine et de faire des liens. Ce chapitre se concentre sur la compréhension des concepts-clés liées aux tests.

! Objectifs d'apprentissage

- Comprendre le rôle de l'incertitude dans la prise de décision.
- Comprendre l'importance du rapport signal/bruit en tant que preuve.
- Connaître les ingrédients de base des tests d'hypothèse et être capable de formuler et d'identifier correctement ces composants dans un article scientifique
- Interpréter correctement les valeurs-p et les intervalles de confiance pour un paramètre.

Avant d'entamer une collecte de données pour une expérience, il est nécessaire de formuler une question de recherche. En général, cette hypothèse spécifie les différences potentielles entre les caractéristiques de la population dues à une intervention (un traitement) que le chercheur souhaite quantifier. C'est à cette étape que les chercheurs décident de la taille de l'échantillon, du choix de la variable de réponse et de la méthode de mesure, qu'ils rédigent le plan de l'étude, etc.

Il est important de noter que la plupart des questions de recherche ne peuvent être résolues à l'aide d'outils simples. Les chercheurs qui souhaitent mener une recherche mé-

thodologique innovante devraient contacter des experts et consulter des statisticien(ne)s **avant** de collecter leurs données afin d'obtenir des informations sur la meilleure façon de procéder pour ce qu'ils ont en tête, afin d'éviter le risque d'affirmations trompeuses basées sur une analyse ou une collecte de données incorrectes.

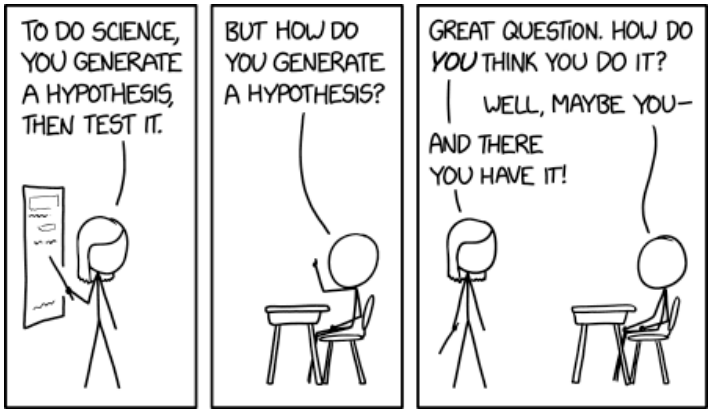

Figure 2.1: Bande dessinée xkcd 2569 (Hypothesis generation) par Randall Munroe. Texte alternatif: Frazzled scientists are requesting that everyone please stop generating hypotheses for a little bit while they work through the backlog. Bande réimprimée sous license CC BY-NC 2.5.

2.1 Variabilité échantillonale

Un chercheur s'intéressera à l'estimation de certaines caractéristiques de la population à partir d'une base de données. Nous pouvons caractériser l'ensemble de toutes les valeurs potentielles que leurs mesures peuvent prendre, ainsi que leur fréquence, au moyen d'une loi d'une variable aléatoire.

L'objectif de cette section est d'illustrer le fait que nous ne pouvons pas simplement utiliser les différences brutes entre les groupes pour effectuer des comparaisons significatives: en raison de la variabilité due à l'échantillonnage, les échantillons seront semblables même s'ils sont générés de la même manière, mais il y aura toujours des différences entre les statistiques récapitulatives calculées sur des échantillons différents. Ces différences ont tendance à s'atténuer (ou à augmenter) au fur et à mesure que l'on collecte davantage d'observations. Plus nous recueillons de données (et donc d'informations) sur notre cible, plus le portrait devient précis. C'est somme toute ce qui nous permet de tirer des conclusions mais, pour ce faire, nous devons d'abord déterminer ce qui est probable ou plausible et donc le fruit du hsard, de ce qui n'est pas ou peu susceptible de se produire.

Nous appelons **statistiques** les résumés numériques des données. Il est important de faire la distinction entre les procédures ou formules et leurs valeurs numériques. Un **estimateur** est une règle ou une formule utilisée pour calculer une estimation d'un paramètre ou d'une quantité d'intérêt sur la base de données observées (comme une recette de gâteau). Une fois que nous disposons de données observées, nous pouvons calculer la moyenne de l'échantillon, c'est-à-dire que nous disposons d'une estimation — d'une valeur réelle (le gâteau), qui est une réalisation unique et non aléatoire. En d'autres termes,

- un estimand est notre cible conceptuelle, comme la caractéristique de la population qui nous intéresse (la moyenne de la population).
- un estimateur est la procédure ou la formule qui nous indique comment transformer les données de l'échantillon en un résumé numérique qui est une approximation de notre cible.
- une estimation (ou un estimé) est un nombre, la valeur numérique obtenue lorsque nous appliquons la formule à un échantillon en praticulier.

(a) Estimand

(b) Estimateur

(c) Estimé

Figure 2.2: Les concepts d'estimand (gauche), estimateur (milieu) et estimaté (droite), illustrés à l'aide de gâteau, une variation d'un idée originale de Simon Grund. Les photos de gâteau sont partagées sous licence CC BY-NC 2.0.

Par exemple, si l'estimand est l'espérance de la population μ, l'estimateur sera la moyenne arithmétique, soit la somme des éléments de l'échantillon aléatoire divisé par la taille de l'échantillon, ou, $\overline{Y} = (Y_1 + \cdots + Y_n)/n$. L'estimé sera une valeur numérique, disons 4.3.

Parce que les intrants de l'estimateur sont aléatoires, la sortie l'est également et varie d'un échantillon à l'autre. Autrement dit, même si on répète une recette, on n'obtient pas le même résultat à chaque coup, comme le montre si bien la Figure 2.3.

Pour illustrer ce point, Figure 2.4 montre cinq échantillons aléatoires simples de taille $n = 10$ tirés d'une population hypothétique de moyenne théorique μ et d'écart-type σ, ainsi que leur moyenne d'échantillon $\overline{y}$. En raison de la variabilité échantillonnale, les

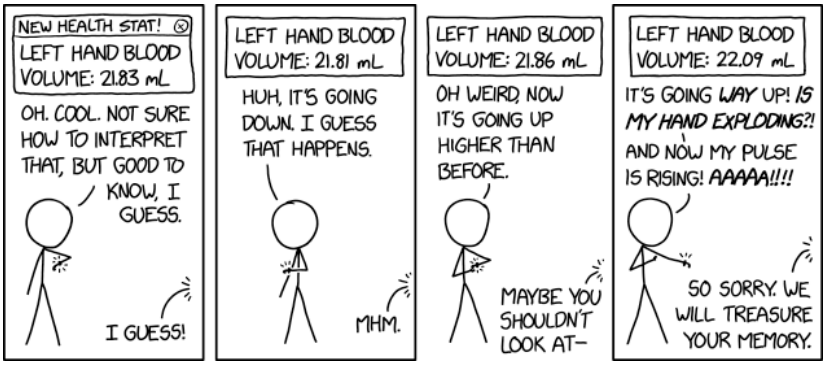

Figure 2.3: Bande dessinée xkcd 2581 (Health Stats) par Randall Munroe. Texte alternatif: You will live on forever in our hearts, pushing a little extra blood toward our left hands now and then to give them a squeeze. Bande réimprimée sous license CC BY-NC 2.5.

moyennes des sous-groupes sont différentes même si elles proviennent de la même population. Vous pouvez considérer la variabilité d'échantillonnage comme du bruit: notre objectif est d'extraire le signal (typiquement les différences de moyennes) tout en tenant compte du bruit de fond.

L'oeil avisé pourra remarquer que les moyennes des cinq échantillons (segments horizontaux colorés) sont moins dispersées autour de la ligne horizontale noire représentant la moyenne de la population μ que ne le sont les observations. Il s'agit là d'un principe fondamental de la statistique: l'information s'accumule au fur et à mesure que l'on obtient plus de données.

Les valeurs de la moyenne de l'échantillon ne donnent pas une image complète et l'étude des différences de moyenne (entre les groupes ou par rapport à une valeur de référence postulée) n'est pas suffisante pour tirer des conclusions. Dans la plupart des cas, rien ne garantit que la moyenne de l'échantillon sera égale à sa valeur réelle, car elle varie d'un échantillon à l'autre: la seule garantie que nous ayons est qu'elle sera en moyenne égale à la moyenne de la population dans des échantillons répétés. Selon le choix de la mesure et la variabilité de la population, il peut y avoir des différences considérables d'une observation à l'autre, ce qui signifie que la différence observée peut être un coup de chance.

Pour avoir une idée du degré de certitude d'une chose, nous devons considérer la variabilité d'une observation Y_i. Cette variance d'une observation tirée de la population est typiquement notée σ^2 et sa racine carrée, l'écart-type, par σ.

L'écart-type *d'une statistique* est appelé **erreur-type**; il ne doit pas être confondu avec l'écart-type σ de la population dont sont tirées les observations de l'échantillon $Y_1, \ldots, Y_n$. L'écart-type et l'erreur-type sont exprimés dans les mêmes unités que les données et sont

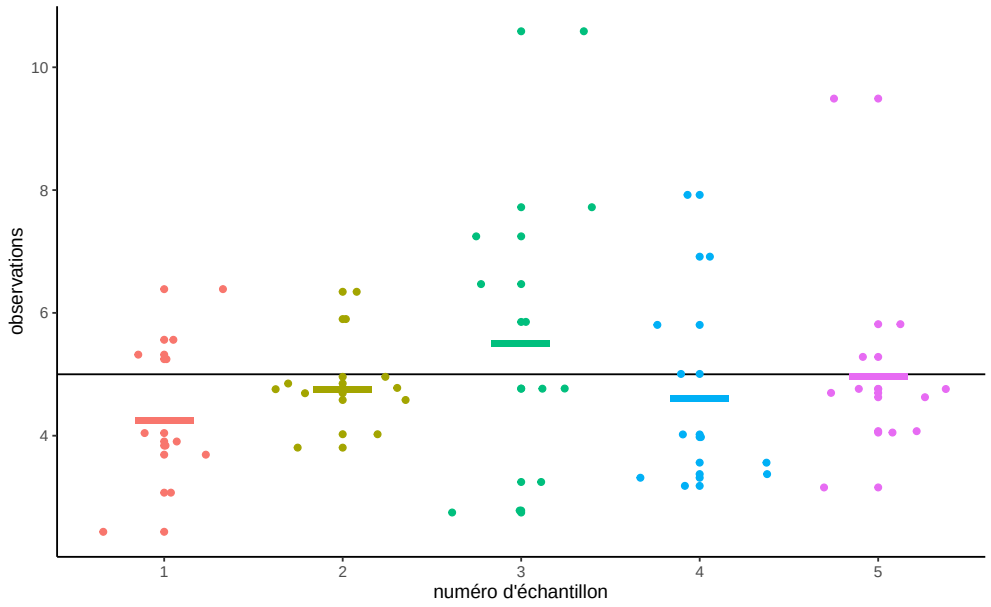

Figure 2.4: Cinq échantillons de taille $n = 10$ tirés d'une population commune de moyenne μ (ligne horizontale). Les segments colorés représentent les moyennes empiriques de chaque groupe.

donc plus faciles à interpréter que la variance. L'erreur-type étant fonction de la taille de l'échantillon, il est d'usage de rapporter plutôt l'écart-type dans les rapports.

Exemple 2.1 (Proportion échantillonale et tirages uniformes)**.** Pour illustrer le concept de variabilité échantillonnale, nous suivons l'exemple de [Matthew Crump] (https://www.crumplab.com/statistics/foundations-for-inference.html) et considérons des échantillons provenant d'une distribution uniforme sur $\{1, 2, \ldots, 10\}$: chaque entier de cet intervalle a la même probabilité d'être tiré.

Même s'ils sont tirés de la même population, les 10 échantillons de Figure 2.5 sont très différents. La seule chose en jeu ici est la variabilité de l'échantillon: puisqu'il y a $n = 20$ d'observations au total, il devrait y avoir en moyenne 10% des observations dans chacun des 10 bacs, mais certains bacs sont vides et d'autres ont plus d'effectifs que prévu. Cette fluctuation est le fruit du hasard.

Comment pouvons-nous donc déterminer si ce que nous voyons est compatible avec le modèle qui, selon nous, a généré les données ? Il suffit de collecter davantage d'observations: la hauteur de la barre est la proportion de l'échantillon, une moyenne de valeurs 0/1, où la valeur 'un' indique que l'observation se trouve dans la case, et 'zéro' dans le cas contraire.

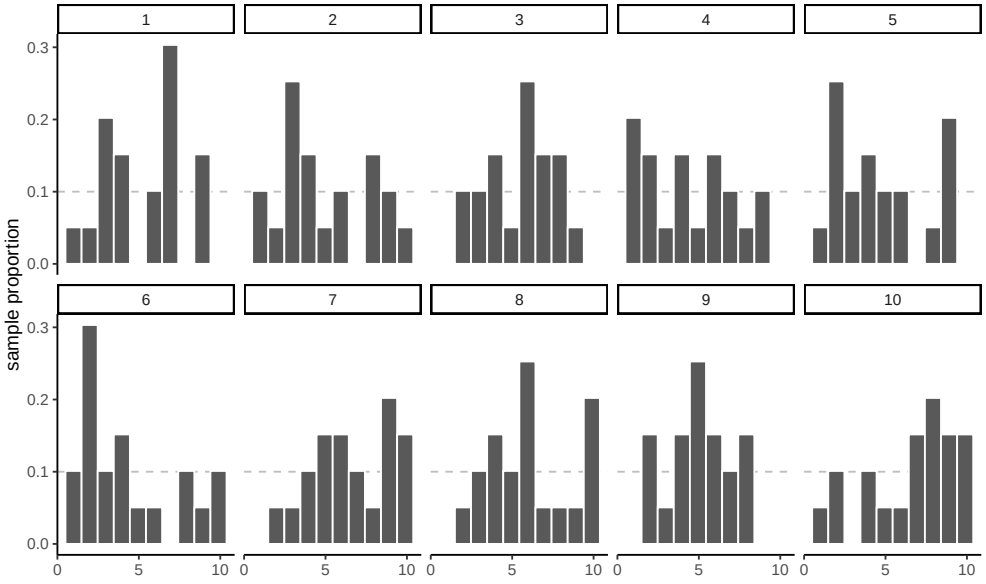

Figure 2.5: Histogrammes de 10 échantillons aléatoires de taille $n = 20$ de loi uniforme discrète.

Considérons maintenant ce qui se passe lorsque nous augmentons la taille de l'échantillon: le panneau supérieur de Figure 2.6 montre des échantillons uniformes pour une taille d'échantillon croissante. Le diagramme à bande ressemble de plus en plus à la véritable distribution sous-jacente (fonction de masse constante, donc chaque case ayant la même fréquence) à mesure que la taille de l'échantillon augmente. La distribution des points de l'échantillon est presque indiscernable de la distribution théorique (ligne droite) lorsque $n = 10000$.[1]. Le panneau du bas, en revanche, ne provient pas d'une distribution uniforme. Plus l'échantillon grossit, plus l'approximation de la fonction de masse se rapproche de la vraie valeur. Nous n'aurions pas pu remarquer cette différence dans les deux premiers graphiques, car la variabilité de l'échantillonnage est trop importante; là, le manque de données dans certaines cases pourrait être un obstacle à l'obtention d'une distribution uniforme.

[1] La formule montre que l'erreur standard diminue d'un facteur 10 chaque fois que la taille de l'échantillon augmente d'un facteur 100.

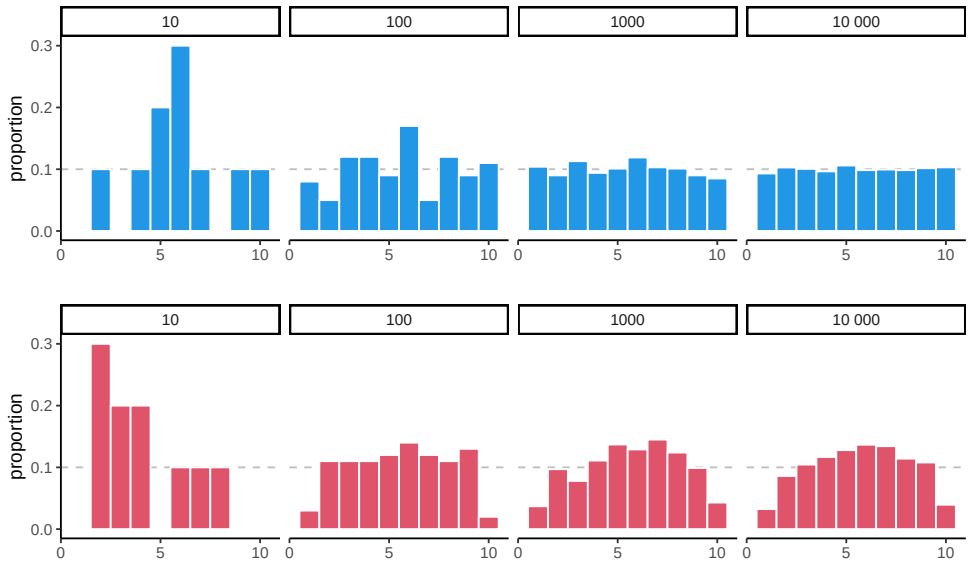

Figure 2.6: Histogrammes de données tirées d'une loi uniforme (haut) et d'une loi non-uniforme (bas) pour des tailles d'échantillons de 10, 100, 1000 and 10 000 (de gauche à droite).

2.2 Tests d'hypothèse

Un test d'hypothèse statistique est une façon d'évaluer la preuve statistique provenant d'un échantillon afin de faire une décision quant à la population sous-jacente. Les étapes principales sont:

- définir les paramètres du modèle,
- formuler les hypothèses alternative et nulle,
- choisir et calculer la statistique de test,
- déterminer son comportement sous $\mathscr{H}_0$ (loi nulle),
- calculer la valeur-p,
- conclure dans le contexte du problème (rejeter ou ne pas rejeter $\mathscr{H}_0$).

Mon approche privilégiée pour présenter les tests d'hypothèse est de faire un parallèle avec un procès pour meurtre où vous êtes nommé juré.

- Le juge vous demande de choisir entre deux hypothèses mutuellement exclusives, coupable ou non-coupable, sur la base des preuves présentées.

- Votre postulat de départ repose sur la présomption d'innocence: vous condamnerez uniquement le suspect si la preuve est accablante. Cela permet d'éviter les erreurs judiciaires. L'hypothèse nulle $\mathscr{H}_0$ est donc *non-coupable*, et l'hypothèse alternative $\mathscr{H}_a$ est coupable. En cas de doute raisonnable, vous émettrez un verdict de non-culpabilité.
- La choix de la statistique de test représente la preuve. Plus la preuve est accablante, plus grande est la chance d'un verdict de culpabilité — le procureur a donc tout intérêt à bien choisir les faits présentés en cour. Le choix de la statistique devrait donc idéalement maximiser la preuve pour appuyer le postulat de culpabilité le mieux possible (ce choix reflète la **puissance** du test).
- En qualité de juré, vous analysez la preuve à partir de la jurisprudence et de l'avis d'expert pour vous assurer que les faits ne relèvent pas du hasard. Pour le test d'hypothèse, ce rôle est tenu par la loi sous $\mathscr{H}_0$: si la personne était innocente, est-ce que les preuves présentées tiendraient la route? des traces d'ADN auront davantage de poids que des ouï-dire (la pièce de théâtre *Douze hommes en colère* de Reginald Rose présente un bel exemple de procès où un des juré émet un doute raisonnable et convainc un à un les autres membres du jury de prononcer un verdict de non-culpabilité).
- Vous émettez un verdict, à savoir une décision binaire, où l'accusé est déclaré soit non-coupable, soit coupable. Si vous avez une valeur-*p*, disons P, pour votre statistique de test et que vous effectuez ce dernier à niveau α, la règle de décision revient à rejeter $\mathscr{H}_0$ si $P < \alpha$.

On s'attarde davantage sur ces définitions heuristiques et le vocabulaire employé pour parler de tests d'hypothèse.

2.3 Hypothèse

Dans les test statistique il y a toujours deux hypothèse: l'hypothèse nulle ($\mathscr{H}_0$) et l'hypothèse alternative ($\mathscr{H}_a$). Habituellement, l'hypothèse nulle est le « statu quo » et l'alternative est l'hypothèse que l'on cherche à démontrer. On se fait l'avocat du Diable en défendant l'hypothèse nulle et en analysant toutes les preuves sous l'angle: « est-ce que les données entrent en contradiction avec $\mathscr{H}_0$? ». Un test d'hypothèse statistique nous permet de décider si nos données nous fournissent assez de preuves pour rejeter $\mathscr{H}_0$ en faveur de $\mathscr{H}_a$, selon un risque d'erreur spécifié.

Généralement, les tests d'hypothèses sont exprimés en fonction de paramètres (de valeurs inconnues) du modèle sous-jacent, par ex. θ. Un test d'hypothèse bilatéral concernant un

paramètre scalaire θ s'exprimerait la forme suivante:

$$\mathscr{H}_0 : \theta = \theta_0 \qquad \text{versus} \qquad \mathscr{H}_a : \theta \neq \theta_0.$$

Ces hypothèses permettent de tester si θ est égal à une valeur numérique précise θ_0.

Par exemple, pour un test bilatéral concernant le paramètre d'un modèle de régression β_j associé à une variable explicative d'intérêt X_j, les hypothèses sont

$$\mathscr{H}_0 : \beta_j = \beta_j^0 \qquad \text{versus} \qquad \mathscr{H}_a : \beta_j \neq \beta_j^0,$$

où β_j^0 est une valeur précise qui est reliée à la question de recherche. Par exemple, si $\beta_j^0 = 0$ la question de recherche sous-jacente est: est-ce que la covariable X_j impacte la variable réponse d'intérêt Y une fois l'effet des autres variables pris en compte?

Il est possible d'imposer une direction dans les tests en considérant une hypothèse alternative de la forme $\mathscr{H}_a : \theta > \theta_0$ ou $\mathscr{H}_a : \theta < \theta_0$.

2.4 Statistique de test

Une statistique de test T est une fonction des données qui résume l'information contenue dans les données pour θ. La forme de la statistique de test est choisie de façon à ce que son comportement sous $\mathscr{H}_0$, c'est-à-dire l'ensemble des valeurs que prend T si $\mathscr{H}_0$ est vraie et leur probabilité relative, soit connu. En effet, T est une variable aléatoire et sa valeur va changer selon l'échantillon. La **loi nulle** de la statistique de test nous permet de déterminer quelles valeurs de T sont plausibles si $\mathscr{H}_0$ est vraie. Plusieurs statistiques que l'on couvrira dans ce cours sont des **statistiques de Wald**, de la forme

$$T = \frac{\widehat{\theta} - \theta_0}{\mathrm{se}(\widehat{\theta})}$$

où $\widehat{\theta}$ est l'estimateur du paramètre θ, θ_0 la valeur numérique postulée (par ex., zéro) et $\mathrm{se}(\widehat{\theta})$ est l'estimateur de l'écart-type de $\widehat{\theta}$.

Par exemple, pour une hypothèse sur la moyenne d'une population de la forme

$$\mathscr{H}_0 : \mu = 0, \qquad \mathscr{H}_a : \mu \neq 0,$$

la statistique de test de Wald est

$$T = \frac{\overline{X} - 0}{S_n/\sqrt{n}}$$

où $\overline{X}$ est la moyenne de l'échantillon $X_1, \ldots, X_n$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

et l'erreur-type de la moyenne $\overline{X}$ est $S_n/\sqrt{n}$; l'écart-type S_n est un estimateur de σ, où

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

2.5 Loi nulle et valeur-*p*

La **valeur-*p*** nous permet de déterminer si la valeur observée de la statistique de test T est plausible sous $\mathscr{H}_0$. Plus précisément, la valeur-*p* est la probabilité, si $\mathscr{H}_0$ est vraie, que la statistique de test soit égale or plus extrême à ce qu'on observe. Supposons qu'on a un échantillon $X_1, \ldots, X_n$ et qu'on observe une valeur de la statistique de test de $T = t$. Pour un test d'hypothèse bilatéral $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_a : \theta \neq \theta_0$, la valeur-*p* est $\Pr_0(|T| \geq |t|)$. Si la distribution de T est symétrique autour de zéro, la valeur-*p* vaut

$$p = 2 \times \Pr_0(T \geq |t|).$$

La Figure 2.7 montre la loi des valeurs-p sous deux scénarios: à gauche, une loi nulle et à droite, une loi alternative. La probabilité de rejetter $\mathscr{H}_0$ est obtenue en calculant l'aire sous la courbe sous la courbe de densité et $\alpha = 0.1$. Sous l'hypothèse nulle, le modèle est calibré et la loi des valeurs-p est uniforme (un rectangle de hauteur 1), ce qui veut dire que toutes les valeurs sont également plausibles. Sous l'alternative, l'obtention de petites valeurs−p est plus plausible.

Il existe généralement trois façons d'obtenir des lois nulles pour évaluer le degré de preuve contre l'hypothèse nulle

- les calculs exacts (combinatoires)
- la théorie des grands échantillons (appelée « régime asymptotique » dans le jargon statistique)
- les méthodes de simulation Monte Carlo.

Bien que souhaitable, la première méthode n'est applicable que dans des cas simples (comme le calcul de la probabilité d'obtenir deux six en lançant deux dés identiques). La deuxième méthode est la plus couramment utilisée en raison de sa généralité et de sa facilité d'utilisation (en particulier dans les temps anciens où la puissance de calcul était rare), mais elle ne donne pas de bons résultats avec des échantillons de petite taille (où

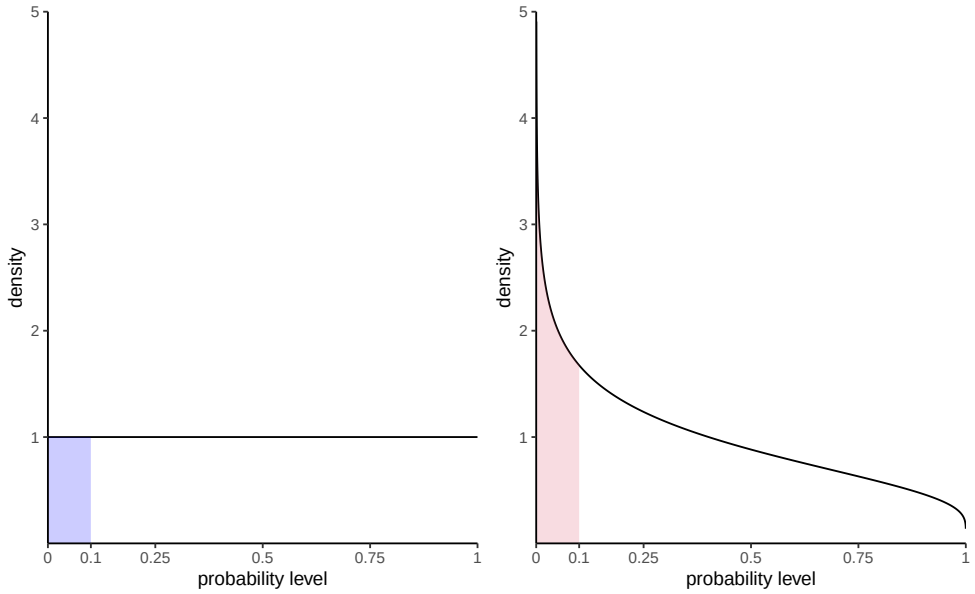

Figure 2.7: Densité des valeurs-p sou l'hypothèse nulle (gauche) et une alternative avec un ratio signal-bruit de 0.5 (droite).

la noti de « trop petit » dépend du contexte et du test). La dernière approche peut être utilisée pour approcher la distribution nulle dans de nombreux scénarios, mais elle ajoute une couche d'aléatoire et les coûts de calcul supplémentaires n'en valent parfois pas la peine.

Prenons l'exemple d'un test d'hypothèse bilatéral pour la moyenne au population $\mathscr{H}_0$: $\mu = 0$ contre $\mathscr{H}_a$: $\mu \neq 0$. Si l'échantillon provient d'une (population de) loi normale $\mathsf{normale}(\mu, \sigma^2)$, on peut démontrer que, si $\mathscr{H}_0$ est vraie et donc $\mu = 0$), la statistique de test

$$T = \frac{\overline{X}}{S/\sqrt{n}}$$

suit une loi de Student-t avec $n-1$ degrés de liberté, dénotée $\mathsf{Student}_{n-1}$. À partir de cette loi nulle, on peut calculer la valeur-p (ou bien à partir d'une table ou d'un logiciel statistique). Puisque la distribution Student-t est symétrique autour de 0, on peut calculer la valeur-p comme $P = 2 \times \Pr(T > |t|)$, où $T \sim \mathsf{Student}_{n-1}$.

2.6 Intervalle de confiance

Un **intervalle de confiance** est une manière alternative de rapporter les conclusions d'un test, en ce sens qu'on fournit une estimation ponctuelle de $\hat{\theta}$ avec une marge d'erreur. L'intervalle de confiance donne donc une indication de la variabilité de la procédure d'estimation. Un intervalle de confiance de Wald à $(1-\alpha)$ pour un paramètre θ est de la forme

$$[\widehat{\theta} + \mathfrak{q}_{\alpha/2}\mathrm{se}(\widehat{\theta}), \widehat{\theta} + \mathfrak{q}_{1-\alpha/2} \times \mathrm{se}(\widehat{\theta})]$$

où $\mathfrak{q}_\alpha$ dénote le quantile d'ordre $\alpha \in (0,1)$ de la loi nulle de la statistique de Wald,

$$T = \frac{\widehat{\theta} - \theta}{\mathrm{se}(\widehat{\theta})},$$

et où θ représente la valeur du paramètre θ (supposé fixe, mais inconnu) de la population.

Par exemple, pour un échantillon aléatoire $X_1, \ldots, X_n$ provenant d'une loi $\mathsf{normale}(\mu, \sigma)$, l'intervalle de confiance à $(1-\alpha)$ pour la moyenne (dans la population) μ est

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

où $t_{n-1,\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi Student-t avec $n-1$ degrés de libertés.

Les bornes de l'intervalle de confiance sont aléatoires puisque $\widehat{\theta}$ et $\mathrm{se}(\widehat{\theta})$ sont des variable aléatoires: leurs valeurs observées changent d'un échantillon à un autre. Avant qu'on calcule l'intervalle de confiance, il y a une probabilité de $1-\alpha$ que θ soit contenu dans l'intervalle **aléatoire** symmétrique $(\widehat{\theta} - \mathfrak{q}_{\alpha/2}\ \mathrm{se}(\widehat{\theta}), \widehat{\theta} + \mathfrak{q}_{\alpha/2}\ \mathrm{se}(\widehat{\theta}))$, où $\widehat{\theta}$ dénote l'estimateur de θ. Une fois qu'on obtient un échantillon et qu'on calcule les bornes de l'intervalle de confiance, il n'y a plus de notion de probabilité: la vraie valeur du paramètre θ (inconnue) est soit contenue dans l'intervalle de confiance, soit pas. La seule interprétation de l'intervalle de confiance qui soit valable alors est la suivante: si on répète l'expérience plusieurs fois et qu'à chaque fois on calcule un intervalle de confiance à $1-\alpha$, alors une proportion de $(1-\alpha)$ de ces intervalles devraient contenir la vraie valeur de θ (de la même manière, si vous lancez une pièce de monnaie équilibrée, vous devriez obtenir grosso modo une fréquence de 50% de pile et 50% de face, mais chaque lancer donnera un ou l'autre de ces choix). Notre « confiance » est dans la procédure et non pas dans les valeurs numériques obtenues pour un échantillon donné.

Si on s'intéresse seulement à la décision rejeter/ne pas rejeter $\mathscr{H}_0$, l'intervalle de confiance est équivalent à la valeur-p en ce sens qu'il mène à la même décision. L'intervalle de confiance donne en revanche l'ensemble des valeurs pour lesquelles la statistique de test

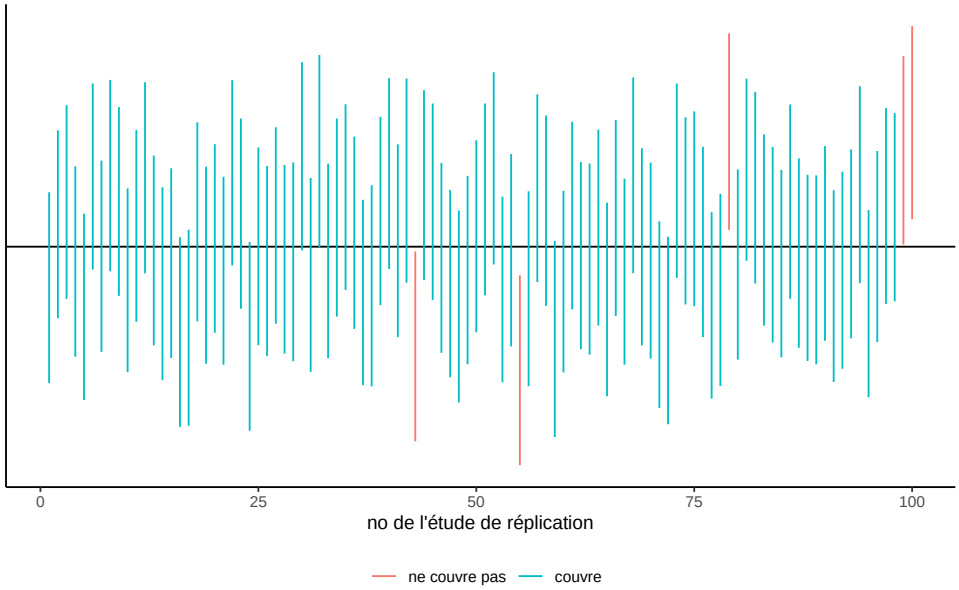

Figure 2.8: Intervalles de confiance à 95% pour la moyenne d'une population normale standard pour 100 échantillons aléatoires. En moyenne, 5% de ces intervalles (en rouge) n'incluent pas la vraie valeur de la moyenne de zéro.

ne fournit pas assez de preuves pour rejeter $\mathscr{H}_0$: pour un test à niveau α, on ne rejetterait aucune des valeurs contenues dans l'intervalle de confiance de niveau $1-\alpha$. Si la valeur-p est inférieure à α, la valeur postulée pour θ est donc hors de l'intervalle de confiance calculé. À l'inverse, la valeur-p ne donne la probabilité d'obtenir un résultat aussi extrême sous l'hypothèse nulle que pour une seule valeur numérique, mais permet de quantifier précisément à quel point le résultat est extrême.

2.7 Conclusion

La valeur-p nous permet de faire une décision quant aux hypothèses du test. Si $\mathscr{H}_0$ est vraie, la valeur-p suit une loi uniforme. Si la valeur-p est petite, ça veut dire que le fait d'observer une statistique de test égal ou encore plus extrême que $T = t$ est peu probable, et donc nous aurons tendance de croire que $\mathscr{H}_0$ n'est pas vraie. Il y a pourtant toujours un risque sous-jacent de commettre un erreur quand on prend une décision. En statistique, il y a deux types d'erreurs:

- erreur de type I: on rejette $\mathscr{H}_0$ alors que $\mathscr{H}_0$ est vraie

- erreur de type II: on ne rejette pas $\mathscr{H}_0$ alors que $\mathscr{H}_0$ est fausse

Ces deux erreurs ne sont pas égales: on cherche souvent à contrôler l'erreur de type I (une erreur judiciaire, condamner un innocent). Pour se prémunir face à ce risque, on fixe préalablement un niveau de tolérance. Plus notre seuil de tolérance α est grand, plus on rejette souvent l'hypothèse nulle même si cette dernière est vraie. La valeur de $\alpha \in (0,1)$ est la probabilité qu'on rejette $\mathscr{H}_0$ quand $\mathscr{H}_0$ est en fait vraie.

$$\alpha = \Pr_0(\text{ rejeter } \mathscr{H}_0).$$

Comme chercheur, on choisit ce niveau α; habituellement 1%, 5% ou 10%. La probabilité de commettre une erreur de type I est α seulement si le modèle nul postulé pour $\mathscr{H}_0$ est correctement spécifié (sic) et correspond au modèle générateur des données.

Le choix du statu quo (typiquement $\mathscr{H}_0$) s'explique plus facilement avec un exemple médical. Si vous voulez prouver qu'un nouveau traitement est meilleur que l'actuel (ou l'absence de traitement), vous devez démontrer hors de tout doute raisonnable que ce dernier ne cause pas de torts aux patients et offre une nette amélioration (pensez à Didier Raoult et ses allégations non-étayées voulant que l'hydrochloroquine, un antipaludique, soit efficace face au virus de la Covid19).

Décision \ vrai modèle	$\mathscr{H}_0$	$\mathscr{H}_a$
ne pas rejeter $\mathscr{H}_0$	✓	erreur de type II
rejeter $\mathscr{H}_0$	erreur de type I	✓

Pour prendre une décision, on doit comparer la valeur-*p* P avec le niveau du test α:

- si $P < \alpha$ on rejette $\mathscr{H}_0$,
- si $P \geq \alpha$ on ne rejette pas $\mathscr{H}_0$.

Attention à ne pas confondre niveau du test (probabilité fixée au préalable par l'expérimentateur) et la valeur-*p* (qui dépend de l'échantillon). Si vous faites un test à un niveau 5% la probabilité de faire une erreur de type I est de 5% par définition, quelque soit la valeur de la valeur-*p*. La valeur-*p* s'interprète comme la probabilité d'obtenir une valeur de la statistique de test égale ou même plus grande que celle qu'on a observée dans l'échantillon, si $\mathscr{H}_0$ est vraie.

Mise en garde

L'*American Statistical Association (ASA)* a publié une liste de principes détaillant les principales erreurs d'interprétation des valeurs-p, notamment

(2) Les valeurs-p ne mesurent pas la probabilité que l'hypothèse étudiée est vrai

(3) Les décisions d'affaires et scientiques ne devraient pas seulement être basées sur le fait qu'une valeur-p est inférieure à un seuil spécifié.

(4) Les analyses statistiques et les valeurs-p associées ne devraient pas être rapportées de manière sélective.

(5) Les valeurs-p, ou la significativité statistiques, ne mesurent pas la taille de l'effet ou l'importance d'un résultat.

2.8 Puissance statistique

Le but du test d'hypothèse est de prouver (hors de tout doute raisonnable) qu'une différence ou un effet est significatif: par exemple, si une nouvelle configuration d'un site web (hypothèse alternative) permet d'augmenter les ventes par rapport au statu quo. Notre capacité à détecter cette amélioration dépend de la puissance du test: plus cette dernière est élevée, plus grande est notre capacité à rejeter $\mathscr{H}_0$ quand ce dernier est faux.

Quand on ne rejette pas $\mathscr{H}_0$ et que $\mathscr{H}_a$ est en fait vraie, on commet une erreur de type II: cette dernière survient avec probabilité $1-\gamma$. La **puissance statistique** d'un test est la probabilité que le test rejette $\mathscr{H}_0$ alors que $\mathscr{H}_0$ est fausse, soit

$$\gamma = \Pr_a(\text{rejeter } \mathscr{H}_0)$$

Selon le choix de l'alternative, il est plus ou moins facile de rejeter l'hypothèse nulle en faveur de l'alternative.

On veut qu'un test ait une puissance élevée, c'est-à-dire, le plus près de 1 possible. Minimalement, la puissance du test devrait être α si on rejette l'hypothèse nulle une fraction α du temps quand cette dernière est vraie. La puissance dépend de plusieurs critères, à savoir:

- la taille de l'effet: plus la différence est grande entre la valeur postulée θ_0 du paramètre sous $\mathscr{H}_0$ et le comportement observé, plus il est facile de le détecter (panneau du milieu de Figure 2.9);
- la variabilité: moins les observations sont variables, plus il est facile de déterminer que la différence observée est significative (les grandes différences sont alors moins plausibles, comme l'illustre le panneau de droite de Figure 2.9);

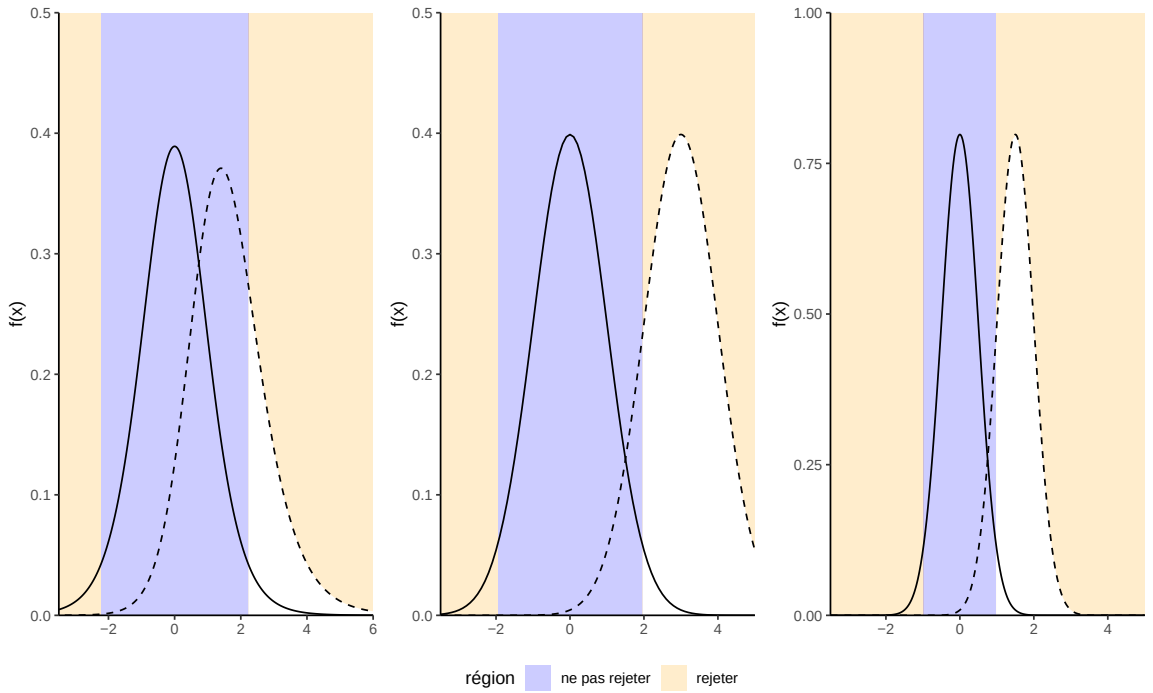

Figure 2.9: Comparaison de la loi nulle (ligne pleine) et d'une alternative spécifique pour un test-t (ligne traitillée). La puissance correspond à l'aire sous la courbe de la densité de la loi alternative qui est dans la zone de rejet du test (en blanc). Le panneau du milieu représente l'augmentation de la puissance suite à l'augmentation de la taille d'effet (différence moyenne entre groupes plus élevée) sous l'hypothèse alternative. Le panneau de droite correspond à un scénario alternatif avec la même taille d'effet, mais une taille d'échantillon ou une précision plus grande.

- la taille de l'échantillon: plus on a d'observations, plus notre capacité à détecter une différence significative augmente parce que l'erreur-type décroît avec la taille de l'échantillon à un rythme (ordinairement) de $n^{-1/2}$. La loi nulle devient aussi plus concentrée quand la taille de l'échantillon augmente.
- le choix de la statistique de test: par exemple, les statistiques basées sur les rangs n'utilisent pas les valeurs numériques qu'à travers le rang relatif. Ces tests sont donc moins puissants parce qu'ils n'utilisent pas toute l'information dans l'échantillon; en contrepartie, ils sont souvent plus robustes en présence de valeurs aberrantes et si le modèle est mal spécifié. Les statistiques de test que nous choisirons sont souvent standards et parmi les plus puissantes qui soient, aussi on ne traitera pas de ce point

davantage dans le cadre du cours.

Pour calculer la puissance d'un test, il faut choisir une alternative spécifique. Pour des exemples simples de statistiques, on peut obtenir une formule explicite pour la puissance. Généralement, on détermine la puissance à l'aide de méthodes de Monte Carlo en simulant des observations d'une alternative donnée, en calculant la statistique de test sur le nouvel échantillon simulé et en calculant la valeur-p associée à notre hypothèse nulle de façon répétée. On calcule par la suite la proportion de tests qui mènent au rejet de l'hypothèse nulle à niveau α, ce qui correspond au pourcentage de valeurs-p inférieures à α.

2.9 Exemples

Exemple 2.2 (Inégalité de genre et tests de permutation)**.** Nous examinons les données de Rosen et Jerdee (1974), qui étudie les stéréotypes de genre et leur impact sur la promotion et les opportunités pour les femmes candidates. L'expérience s'est déroulée en 1972 et les unités expérimentales, composées de 95 superviseurs bancaires masculins, ont reçu divers mémorandums et ont été invitées à fournir des évaluations de candidatures pour un poste de cadre. Ils devaient prendre des décisions sur la base des informations fournies.

Nous nous intéressons à l'expérience 1 relative à la promotion des employés: les responsables devaient décider de promouvoir ou non un employé au poste de directeur de succursale sur la base de recommandations et d'évaluations du potentiel de relations avec les clients et les employés. L'intervention des auteurs s'est concentrée sur la description de la nature (complexité) du travail du gestionnaire (simple ou complexe) et sur le sexe du candidat (homme ou femme): tous les dossiers étaient par ailleurs similaires.

Pour des raisons de simplicité, nous ne considérons que le facteur sexe et nous agrégeons sur le poste pour les $n = 93$ réponses. La table Tableau 2.2 montre le décompte des recommendations pour chaque possibilité.

Tableau 2.2: Recommendations de promotion pour le poste de gestionnaire de branche selon le sexe de la personne qui postule.

	male	female
promouvoir	32	19
ne pas promouvoir	12	30

L'hypothèse nulle qui nous intéresse ici est que le sexe n'a pas d'impact, de sorte que la probabilité de promotion est la même pour les hommes et les femmes. Soit p_{h} et p_{f} ces

probabilités respectives; nous pouvons donc écrire mathématiquement l'hypothèse nulle comme $\mathscr{H}_0 : p_\text{h} = p_\text{f}$ contre l'alternative $\mathscr{H}_a : p_\text{h} \neq p_\text{f}$.

La statistique de test généralement employée pour les tableaux de contingence est un test du chi carré[2], qui compare les proportions globales de promotion de chaque sous-groupe. La proportion de l'échantillon pour les hommes est de 32/42 = ~76%, contre 19/49 =~49% pour les femmes. Bien que cette différence de 16 % semble importante, elle pourrait être trompeuse: l'erreur type pour les proportions de l'échantillon est d'environ 3.2 % pour les hommes et 3.4 % pour les femmes.

S'il n'y avait pas de discrimination fondée sur le sexe, nous nous attendrions à ce que la proportion de personnes promues soit la même dans l'ensemble; elle est de 51/93 ou 0.55 pour l'échantillon regroupé. Nous pourrions nous contenter de tester la différence moyenne, mais nous nous appuyons plutôt sur le test de contingence X_p^2 de Pearson (également appelé test du khi-carré), qui compare les chiffres attendus (sur la base de taux de promotion égaux) aux chiffres observés, convenablement normalisés. convenablement normalisés. Si l'écart est important entre les chiffres attendus et les chiffres observés, cela met en doute la véracité de l'hypothèse nulle.

Si les effectifs de chaque cellule sont importants, la distribution nulle du test du chi-deux est bien approximée par une distribution de χ^2. La sortie du test comprend la valeur de la statistique, 10.79, les degrés de liberté de l'approximation χ^2 et la valeur *p*, qui donne la probabilité qu'un tirage aléatoire d'une distribution χ_1^2 soit plus grand que la statistique de test observée **en supposant que l'hypothèse nulle est vraie**. La valeur *p* est très petite, 0.001, ce qui signifie qu'il est très peu probable qu'un tel résultat soit le fruit du hasard s'il n'y a pas eu de discrimination fondée sur le sexe.

Une autre solution pour obtenir un point de référence permettant d'évaluer le caractère exagéré du rapport de cotes observé consiste à utiliser des simulations: les tests de permutation sont efficaces [illustrés par Jared Wilber] (https://www.jwilber.me/permutationtest/). Considérons une base de données contenant les données brutes avec 93 lignes, une pour chaque gestionnaie, avec pour chacune un indicateur d'`action` et le `sexe` de l'employé hypothétique présenté dans la tâche.

Tableau 2.3: Les cinq premières lignes de la base de données en format long pour l'expérience 1 de Rosen et Jerdee (1974).

action	sexe
promouvoir	homme

[2]Si vous avez suivi des cours de modélisation avancés, il s'agit d'un test de score obtenu en ajustant une régression de Poisson avec `sexe` et `action` comme covariables; l'hypothèse nulle correspondant à l'absence de terme d'interaction entre les deux.

ne pas promouvoir	femme
promouvoir	homme
ne pas promouvoir	femme
ne pas promouvoir	homme

Sous l'hypothèse nulle, le sexe n'a aucune incidence sur l'action du gestionnaire. Cela signifie que nous pourrions dresser un portrait du monde sans discrimination en mélangeant les étiquettes de sexe de manière répétée. Ainsi, nous pourrions obtenir une référence en répétant les étapes suivantes plusieurs fois :

1. permuter les étiquettes pour le `sexe`,
2. recréer un tableau de contingence en agrégeant les effectifs,
3. calculer une statistique de test pour le tableau simulé.

Comme statistique de test, nous utilisons le rapport des cotes: la probabilité d'un événement est le rapport entre le nombre de succès et le nombre d'échecs. Dans notre exemple, il s'agirait du nombre de dossiers promus par rapport au nombre de dossiers retenus. La probabilité de promotion d'un homme est de $32/12$, alors que celle d'une femme est de $19/30$. Le rapport des cotes pour un homme par rapport à une femme est donc $\mathsf{RC} = (32/12)/(19/30) = 4.21$. Sous l'hypothèse nulle, $\mathscr{H}_0 : \mathsf{OR} = 1$ (même probabilité d'être promu) (pourquoi ?)

L'histogramme de la Figure 2.10 montre la distribution du rapport de cotes sur la base de 10 000 permutations. Il est rassurant de constater que nous obtenons à peu près la même valeur *p* approximative, ici 0.002.[3].

L'article concluait (à la lumière de ce qui précède et d'autres expériences)

> Les résultats ont confirmé l'hypothèse selon laquelle les administrateurs masculins ont tendance à discriminer les employées dans les décisions concernant la promotion, le développement et la supervision du personnel.

Récapitulatif

- Paramètres du modèle: probabilité de promotion pour les hommes et les femmes, respectivement p_h et p_f.
- Hypothèses: pas de discrimination fondée sur le sexe, ce qui signifie une probabilité de promotion égale (hypothèse nulle $\mathscr{H}_0 : p_\text{h} = p_\text{f}$, contre hypothèse alternative $\mathscr{H}_a : p_\text{h} \neq p_\text{f}$).

[3]La valeur *p* obtenue pour le test de permutation changerait d'une exécution à l'autre puisque les intrants sont aléatoires. Cependant, la précision de la statistique est suffisante pour la prise de décision

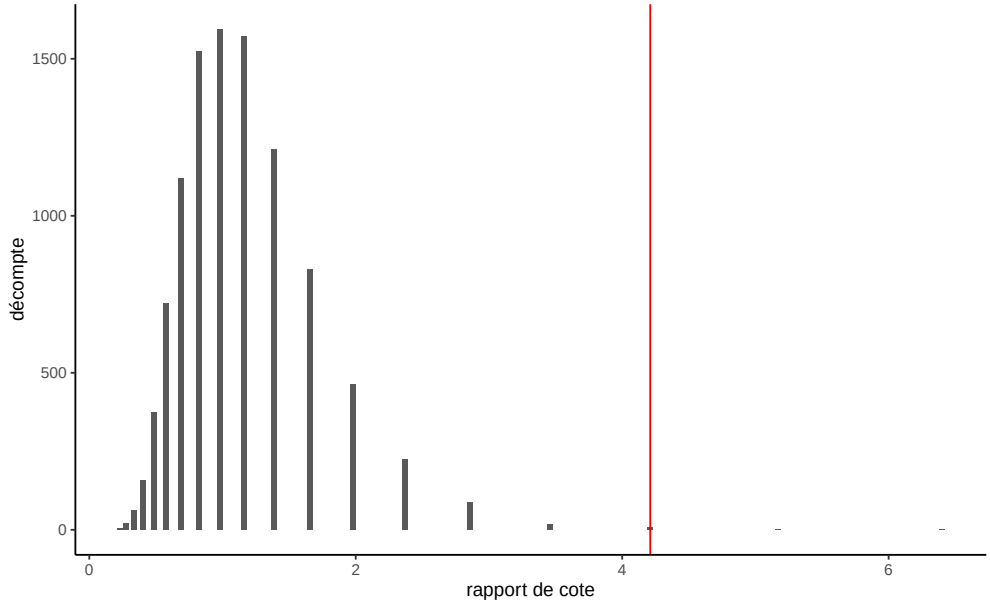

Figure 2.10: Histogramme de simulations de la loi nulle pour le rapport de cote, obtenu par le biais d'un test de permutation; la ligne verticale rouge indique le rapport de cote échantillonal.

- Statistique de test: (1) test du khi-deux pour les tableaux de contingence et (2) rapport de cotes.
- Valeur-p: (1) $.0010$ et (2) $.0024$ pour le test de permutation.
- Conclusion: rejeter l'hypothèse nulle, car il existe des preuves d'une discrimination fondée sur le sexe, avec une probabilité de promotion différente pour les hommes et les femmes.

Conformément aux directives de l'APA, la statistique χ^2 serait présentée sous la forme $\chi^2(1, n = 93) = 10.79$, $p = .001$ en même temps que les effectifs et les proportions de l'échantillon.

Exemple 2.3 (L'élément de surprise d'une prise de contact inattendue)**.** Liu et al. (2023) étudie les interactions sociales et l'impact de la surprise sur les personnes qui contactent de vieilles connaissances de manière inattendue. L'expérience 1 se concentre sur des questionnaires où la condition expérimentale est l'appréciation perçue du fait d'envoyer une communication à quelqu'un avec qui on n'a pas correspondu depuis longtemps (par opposition au fait de se faire contacter). L'étude a utilisé un questionnaire envoyé à 200 adultes américains recrutés sur la plateforme Prolific Academic. L'indice de réponse consiste en la moyenne de quatre questions mesurées sur une échelle de Likert allant de

1 à 7, les valeurs les plus élevées indiquant une plus grande appréciation de la prise de contact.

Nous pouvons commencer par examiner les statistiques sommaires des variables sociodémographiques (sexe et âge) afin d'évaluer si l'échantillon est représentatif de la population générale dans son ensemble. La proportion d'« autres » (comprenant les personnes non binaires) est beaucoup plus élevée que celle du recensement général, et la population est plutôt jeune selon Tableau 2.4.

Tableau 2.4: Statistiques descriptives de l'âge des participants, et décompte par genre.

genre	min	max	moyenne	n
homme	18	78	32.0	105
femme	19	68	36.5	92
autre	24	30	27.7	3

Tableau 2.5: Appréciation moyenne (écart-type), et nombre de participants par condition expérimentale.

rôle	moyenne	écart-type	n
initiateur	5.50	1.28	103
destinataire	5.87	1.27	97

Comme il n'y a que deux groupes sans chevauchements (c'est à dire que les personnes ont un seul rôle), soit initiateur ou destinataire, le test logique à utiliser est un test-t pour deux échantillons indépendants, ou une variante de celui-ci. En utilisant la statistique du t-test de Welch, la moyenne et l'écart-type de chaque groupe sont estimés à l'aide des données fournies.

Le logiciel renvoie comme valeur du test , ce qui conduit au rejet de l'hypothèse nulle d'absence de différence d'appréciation en fonction du rôle de l'individu (initiateur ou destinataire). La différence moyenne estimée est $\Delta M = -0.37$, 95% CI $[-0.73, -0.01]$; puisque 0 n'est pas inclus dans l'intervalle de confiance, nous rejetons également l'hypothèse nulle au niveau 5%. L'estimation suggère que les initiateurs sous-estiment l'importance de contacter de manière inattendue.[4].

Récapitulatif

[4]En supposant que la variance de chaque sous-groupe soit égale, nous aurions pu utiliser un t-test à deux échantillons à la place. La différence dans la conclusion est insignifiante, avec une valeur *p* presque égale

- Paramètres du modèle: score d'appréciation moyen μ_{i} et μ_{d} des initiateurs et des destinataires, respectivement.
- Hypothèse: le score d'appréciation attendu est le même pour les initiateurs et les destinataires, $\mathscr{H}_0 : \mu_{\text{i}} = \mu_{\text{d}}$ contre l'alternative $\mathscr{H}_0 : \mu_{\text{i}} \neq \mu_{\text{r}}$ qu'ils sont différents.
- Statistique de test: test-t de Welch pour deux échantillons indépendants
- Valeur-p: 0.041
- Conclusion: rejet de l'hypothèse nulle, le score moyen d'appréciation diffère selon le rôle tenu.

Exemple 2.4 (Les communications virtuelles réduisent le nombre d'idées créatives)**.** Une étude de Nature a réalisé une expérience pour voir comment les communications virtuelles impactent le travail d'équipe en comparant le nombre d'idées créatives générées par des binômes au cours d'une tempête d'idée, ainsi que leur qualité telle que mesurée par des arbitres externes. L'échantillon était composé de 301 paires de participants qui ont interagi par vidéoconférence ou en face à face.

Les auteurs ont comparé le nombre d'idées créatives, un sous-ensemble d'idées générées avec un score de créativité supérieur à la moyenne. Le nombre moyen d'idées créatives pour le face à face est 7.92 idées (écart-type 3.40), comparativement à 6.73 idées (écart-type, 3.27) pour la vidéoconférence.

Brucks et Levav (2022) a utilisé un modèle de régression binomiale négative: dans leur modèle, le nombre moyen d'idées créatives générées est

$$\mathsf{E}(\texttt{ncreative}) = \exp(\beta_0 + \beta_1 \texttt{video})$$

où $\texttt{video} = 0$ si la paire se trouve dans la même pièce et $\texttt{video} = 1$ si elle interagit plutôt par vidéoconférence.

Le nombre moyen d'idées pour la vidéoconférence est donc $\exp(\beta_1)$ multiplié par celui du face à face: l'estimation du facteur multiplicatif est $\exp(\beta_1)$ est 0.85 95% CI $[0.77, 0.94]$.

L'absence de différence entre les conditions expérimentales se traduit par l'hypothèse nulle $\mathscr{H}_0 : \beta_1 = 0$ vs $\mathscr{H}_0 : \beta_1 \neq 0$ ou, de manière équivalente, $\mathscr{H}_0 : \exp(\beta_1) = 1$. Le test du rapport de vraisemblance comparant le modèle de régression avec et sans `video` la statistique est $R = 9.89$ (valeur-p basée sur χ^2_1 de $.002$). Nous concluons que le nombre moyen d'idées est différent, les statistiques sommaires suggérant que les paires virtuelles génèrent moins d'idées.

Si nous avions eu recours à un test-t pour deux échantillons indépendants, nous aurions trouvé une différence moyenne dans le nombre d'idées créatives de $\Delta M = 1.19$, 95% CI $[0.43, 1.95]$, $t(299) = 3.09$, $p = .002$.

Les deux tests reposent sur des hypothèses légèrement différentes, mais aboutissent à des conclusions similaires: il a de forts indices que le nombre d'idées créatives est plus faible lorsque les personnes interagissent par vidéoconférence.

Exemple 2.5 (Prix de billets de trains à grande vitesse espagnols)**.** La compagnie nationale de chemin de fer Renfe gère les trains régionaux et les trains à haute vitesse dans toute l'Espagne. Les prix des billets vendus par Renfe sont aggrégés par une compagnie. On s'intéresse ici à une seule ligne, Madrid–Barcelone. Notre question scientifique est la suivante: est-ce que le prix des billets pour un aller (une direction) est plus chère pour un retour? Pour ce faire, on considère un échantillon de 10000 billets entre les deux plus grandes villes espagnoles. On s'intéresse au billets de TGV vendus (AVE) au tarif Promotionnel. Notre statistique de test sera simplement la différence de moyenne entre les deux échantillons: la différence entre le prix en euros d'un train Madrid–Barcelone (μ_1) et le prix d'un billet Barcelone–Madrid (μ_2) est $\mu_1 - \mu_2$ et notre hypothèse nulle est qu'il n'y a aucune différence de prix, soit $\mathscr{H}_0 : \mu_1 - \mu_2 = 0$.

On utilise de nouveau le test de Welch pour deux échantillons en filtrant les données pour ne conserver que les billets au tarif Promo: la moyenne des billets Barcelone-Madrid est 82.11 euros, ceux pour Madrid-Barcelone 82.56 euros et la valeur de la statistique de Welch est -1.33. Si on utilise l'approximation normale, on obtient une valeur-p de 0.18.

Plutôt que d'utiliser la loi asymptotique (qui est valide pour de grands échantillons à cause du théorème central limite), on peut considérer une approximation sous une hypothèse moins restrictive en supposant que les données sont échangeables. Sous l'hypothèse nulle, il n'y aucune différence entre les deux destinations et les étiquettes pour la destination (une variable catégorielle binaire) sont arbitraires. On pourrait considérer les mêmes données, mais avec une permutation des variables explicatives: c'est ce qu'on appelle un test de permutation. On va recréer deux groupes de taille identique à notre échantillon original, mais en changeant les observations. On recalcule la statistique de test sur ces nouvelle données (si on a une poignée d'observations, il est possible de lister toutes les permutations possibles; typiquement, il suffit de considérer un grand nombre de telles permutations, disons 9999). Pour chaque nouveau jeu de données, on calculera la statistique de test et on calculera le rang de notre statistique par rapport à cette référence. Si la valeur de notre statistique observée sur l'échantillon original est extrême en comparaison, c'est autant de preuves contre l'hypothèse nulle.

La valeur-*p* du test de permutation, 0.186, est la proportion de statistiques plus extrêmes que celle observée. Cette valeur-*p* est quasi-identique à celle de l'approximation de Satterthwaite, à savoir 0.182 (la loi Student-t est numériquement équivalente à une loi standard normale avec autant de degrés de liberté), tel que représenté dans la Figure 2.11. Malgré que notre échantillon soit très grand, avec $n = 8059$ observations, la différence n'est pas

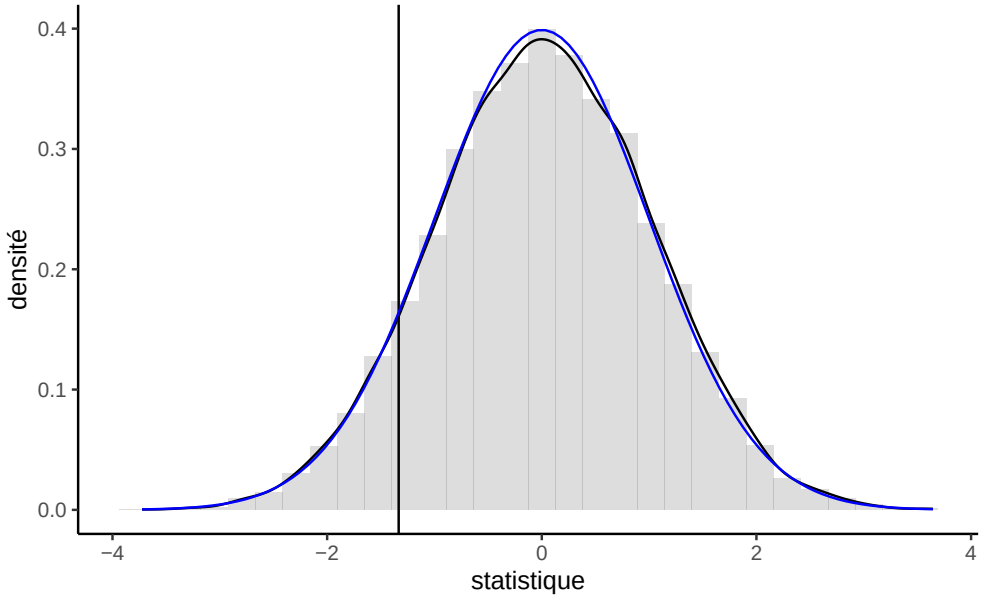

Figure 2.11: Approximation par permutation de la loi nulle de la statistique de test de Welch (histogramme et trait noir) et loi asymptotique normale standard (trait bleu) pour le prix de billets de trains AVE au tarif promotionnel entre Madrid et Barcelone. La valeur de la statistique de test de l'échantillon original est représentée par un trait vertical.

jugée significative. Avec un échantillon de deux millions de billets, on pourrait estimer précisément la moyenne (au centime près): la différence de prix entre les deux destinations et cette dernière deviendrait statistiquement significative. Elle n'est pas en revanche pas pertinente en partique, car une différence de 0.28 euros sur un prix moyen de 82.56 euros est quantité négligeable.

3 Inférence basée sur la vraisemblance

Ce chapitre traite de modélisation statistique et d'inférence basée sur la vraisemblance, la méthodologie la plus populaire dans le monde de la statistique.

> **!** Important
>
> **Objectifs d'apprentissage**
>
> - Apprendre la terminologie associée à l'inférence basée sur la vraisemblance.
> - Dériver des expressions explicites pour l'estimateur du maximum de vraisemblance de modèles simples.
> - En utilisant l'optimisation numérique, obtenir des estimations de paramètres et leurs erreurs-type en utilisant le maximum de vraisemblance.
> - Utiliser les propriétés de la vraisemblance pour les grands échantillons afin d'obtenir des intervalles de confiance et les propriétés des tests statistiques.
> - Être capable d'utiliser les critères d'information pour la sélection des modèles.

Un modèle statistique spécifie typiquement un mécanisme de génération de données. Nous postulons ainsi que les données ont été générées à partir d'une loi de probabilité dotée de p paramètres $\boldsymbol{\theta}$. L'espace d'échantillonnage est l'ensemble dans lequel se trouvent les n observations, tandis que l'espace des paramètres $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ est l'ensemble des valeurs que peuvent prendre le vecteur de paramètres.

Nous considérons un exemple pour motiver les concepts présentés ci-après. Supposons qu'on s'intéresse au temps qu'un usager doit attendre à la station Université de Montréal s'il arrive à 17h59 précise tous les jours de la semaine, juste à temps pour la prochaine rame de métro. La base de données `attente` consistent le temps en secondes avant que la prochaine rame ne quitte la station. Les données ont été collectées pendant trois mois et peuvent être traitées comme un échantillon indépendant. Le panneau gauche de Figure 3.1 montre un histogramme des observations $n = 62$ qui vont de 4 à 57 secondes. Les données sont positives, notre modèle doit donc tenir compte de cette caractéristique.

Exemple 3.1 (Modèle exponentiel pour les temps d'attente)**.** Pour modéliser les temps d'attente, on considère une loi exponentielle avec paramètre d'échelle λ (Définition 1.8),

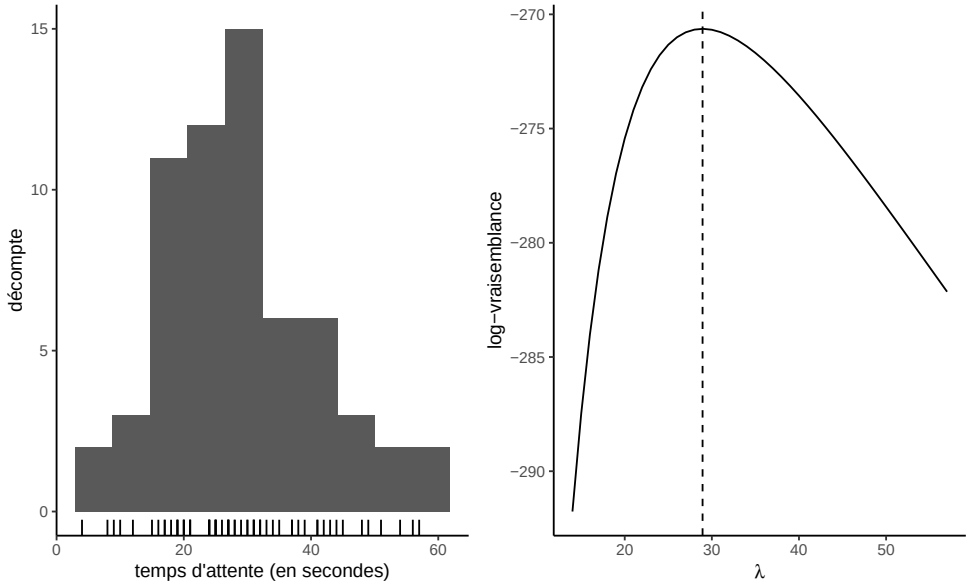

Figure 3.1: Histogramme du temps d'attente avec des traits indiquant les temps observés (gauche) et log-vraisemblance exponentielle, avec la valeur de l'estimation du maximum de vraisemblance en traitillé (droite).

où λ est l'espérance (moyenne théorique). Sous un postulat d'indépendence[1], la densité conjointe des observations $y_1, \ldots, y_n$ est

$$f(\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \lambda^{-1} \exp(-y_i/\lambda) = \lambda^{-n} \exp\left(-\sum_{i=1}^{n} y_i/\lambda\right)$$

L'espace d'échantillonnage est $\mathbb{R}_+^n = [0, \infty)^n$, et l'espace des paramètres est $(0, \infty)$.

Pour estimer le paramètre d'échelle λ et obtenir des mesures d'incertitude appropriées, nous avons besoin d'un **cadre de modélisation**.

3.1 Estimation par maximum de vraisemblance

Pour chaque valeur du paramètre $\boldsymbol{\theta}$, on obtient une fonction de densité ou de masse pour les obserations qui varie en fonction de la compatibilité entre le modèle et les données

[1] Si A et B sont des variables aléatoires indépendantes, leur probabilité conjointe est le produit des probabilités des événements individuels, $\Pr(A \cup B) = \Pr(A)\Pr(B)$. La même factorisation tient pour la fonction de densité ou de masse, lesquelles sont les dérivées de la fonction de répartition.

recueillies. Cela nous permet d'obtenir une fonction objective pour l'estimation des paramètres

Définition 3.1 (Vraisemblance)**.** La **vraisemblance** $L(\boldsymbol{\theta})$ est une fonction des paramètres $\boldsymbol{\theta}$ qui donne la probabilité (ou densité) d'observer un échantillon selon une loi postulée, en traitant les observations comme fixes,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}),$$

où $f(\boldsymbol{y}; \boldsymbol{\theta})$ désigne la densité ou la fonction de masse conjointe du n-vecteur des observations.

Si ces dernières sont indépendantes, la densité conjointe se factorise en un produit de densité unidimensionnelle pour chaque observation et la vraisemblance devient alors

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}) = f_1(y_1; \boldsymbol{\theta}) \times \cdots \times f_n(y_n; \boldsymbol{\theta}).$$

La fonction de log-vraisemblance correspondante pour des données indépendantes et identiquement distribuées est

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta})$$

Exemple 3.2 (Données dépendantes)**.** La fonction de densité conjointe ne se factorise que pour les données indépendantes, mais une décomposition séquentielle alternative peut s'avérer utile. Par exemple, nous pouvons écrire la densité conjointe $f(y_1, \ldots, y_n)$ en utilisant la factorisation

$$f(\boldsymbol{y}) = f(y_1) \times f(y_2 \mid y_1) \times \ldots f(y_n \mid y_1, \ldots, y_n)$$

en termes de densitées conditionnelles. Une telle décomposition est particulièrement utile pour les séries temporelles, où les données sont ordonnées du temps 1 au temps n et où les modèles relient généralement l'observation y_n à son passé. Par exemple, le processus $\mathsf{AR}(1)$ stipule que $Y_t \mid Y_{t-1} = y_{t-1} \sim \mathsf{normale}(\alpha + \beta y_{t-1}, \sigma^2)$ et nous pouvons simplifier la log-vraisemblance en utilisant la propriété de Markov, qui stipule que la réalisation actuelle dépend du passé, $Y_t \mid Y_1, \ldots, Y_{t-1}$, uniquement à travers la valeur la plus récente Y_{t-1}. La log-vraisemblance devient donc

$$\ell(\boldsymbol{\theta}) = \ln f(y_1) + \sum_{i=2}^{n} f(y_i \mid y_{i-1}).$$

Définition 3.2 (Estimateur du maximum de vraisemblance)**.** L'**estimateur du maximum de vraisemblance** (EMV) $\widehat{\boldsymbol{\theta}}$ est la valeur du vecteur qui maximise la vraisemblance,

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \boldsymbol{y}).$$

Le logarithme naturel $\ln$ est une transformation monotone, il est donc préférable de calculer les EMV sur l'échelle logarithmique pour éviter les imprécisions numériques et maximiser de manière équivalente la log-vraisemblance $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ln L(\boldsymbol{\theta}; \boldsymbol{y})$.[2]

Si nous supposons que notre modèle est correct, nous nous attendons à observer ce qui a été réalisé, et nous trouvons donc le vecteur de paramètres qui rend l'échantillon le plus susceptible d'avoir été généré par notre modèle. Plusieurs propriétés de l'estimateur du maximum de vraisemblance le rendent intéressant pour l'inférence. L'estimateur du maximum de vraisemblance est efficace, c'est-à-dire qu'il présente l'erreur quadratique moyenne asymptotique la plus faible de tous les estimateurs. L'estimateur du maximum de vraisemblance est également **convergent**, c'est-à-dire qu'il approche de la vraie valeur du paramètre inconnu à mesure que la taille de l'échantillon augmente (asymptotiquement sans biais).

La plupart du temps, nous allons recourir à des routines d'optimisation numérique pour trouver la valeur de l'estimation du maximum de vraisemblance, ou parfois dériver des expressions explicites pour l'estimateur, à partir de la log-vraisemblance. Le panneau de droite de Figure 3.1 montre la log-vraisemblance exponentielle, qui atteint un maximum à $\widehat{\lambda} = 28.935$ secondes, la moyenne de l'échantillon des observations. La fonction diminue de part et d'autre de ces valeurs à mesure que les données deviennent moins compatibles avec le modèle. Compte tenu de l'échelle pour la log-vraisemblance, ici pour un petit échantillon, il est facile de voir que l'optimisation directe de la fonction de vraisemblance (plutôt que de son logarithme naturel) pourrait conduire à un débordement numérique, puisque $\exp(-270) \approx 5.5 \times 10^{-118}$, et que les valeurs logarithmiques inférieures à -746 seraient arrondies à zéro.

Exemple 3.3 (Calcul de l'estimateur du maximum de vraisemblance d'une loi exponentielle)**.** La Figure 3.1 révèle que la log-vraisemblance exponentielle est unimodale. Nous pouvons utiliser le calcul différentiel pour obtenir une expression explicite pour $\widehat{\lambda}$ sur la base de la log-vraisemblance

$$\ell(\lambda) = -n \ln \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} y_i.$$

[2] Puisque dans la plupart des cas on a un produit de densités, prendre le logarithme transforme un produit de termes potentiellement petits en une somme de log densités, ce qui est plus facile côté dérivation et plus stable du point de vue du calcul numérique.

Si on calcule la dérivée première et que l'on fixe cette dernière à zéro, on obtient

$$\frac{\mathrm{d}\ell(\lambda)}{\mathrm{d}\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} y_i = 0.$$

En réarrangeant cette expression pour amener $-n/\lambda$ à droite de l'égalité, et en multipliant les deux côtés par $\lambda^2 > 0$, on obtient que le point d'inflexion se situe à $\widehat{\lambda} = \sum_{i=1}^{n} y_i/n$. La dérivée deuxième de la log vraisemblance est $\mathrm{d}^2\ell(\lambda)/\mathrm{d}\lambda^2 = n(\lambda^{-2} - 2\lambda^{-3}\overline{y})$, et si on évalue cette dernière à $\lambda = \overline{y}$, on trouve une valeur négative, $-n/\overline{y}^2$. Cela confirme que $\widehat{\lambda}$ est la valeur où la fonction atteint son maximum.

Exemple 3.4 (Échantillons de loi normale)**.** Supposons que nous disposions de n observations de loi normale de paramètres de moyenne μ et de variance σ^2, où $Y_i \sim \mathsf{normale}(\mu, \sigma^2)$ sont indépendants. Rappelons que la densité de la loi normale est

$$f(y; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

Pour une réalisation $y_1, \ldots, y_n$ tirée d'un échantillon aléatoire simple, la vraisemblance est

$$\begin{aligned} L(\mu, \sigma^2; \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}. \end{aligned}$$

et la log-vraisemblance s'écrit

$$\ell(\mu, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

On peut montrer que les estimateurs du maximum de vraisemblance pour les deux paramètres sont

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Le fait que l'estimateur de la moyenne théorique μ soit la moyenne de l'échantillon est assez intuitif et on peut montrer que l'estimateur est sans biais pour μ. L'estimateur sans biais de la variance de l'échantillon est

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Puisque $\widehat{\sigma}^2 = (n-1)/nS^2$, il s'ensuit que l'estimateur du maximum de vraisemblance de σ^2 est biaisé, mais les deux estimateurs sont convergents et s'approcheront donc de la vraie valeur σ^2 pour n suffisamment grand.

Exemple 3.5 (Moindres carrés ordinaires)**.** Le cas des données normalement distribuées est intimement lié à la régression linéaire et aux moindres carrés ordinaires: en supposant la normalité des erreurs, les estimateurs des moindres carrés de $\boldsymbol{\beta}$ coïncident avec l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$.

Le modèle de régression linéaire spécifie que $Y_i \sim \mathsf{normale}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$, ou de manière équivalente

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + \ldots + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad (i = 1, \ldots, n),$$

avec des aléas $\varepsilon_i \sim \mathsf{normale}(0, \sigma^2)$. Le modèle linéaire a $p + 2$ paramètres ($\boldsymbol{\beta}$ et σ^2) et la log-vraisemblance est

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left\{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right\}^2.$$

Maximiser la log-vraisemblance par rapport à $\boldsymbol{\beta}$ équivaut à minimiser la somme des erreurs quadratiques $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2$. Cette fonction objective étant la même que celle des moindres carrés, il s'ensuit que l'estimateur des moindres carrés $\widehat{\boldsymbol{\beta}}$ pour les paramètres de la moyenne est aussi l'estimateur du maximum de vraisemblance si les aléas ont la même variance σ^2, quelle que soit la valeur de cette dernière. L'estimateur du maximum de vraisemblance $\widehat{\sigma}^2$ est donc

$$\widehat{\sigma}^2 = \max_{\sigma^2} \ell(\widehat{\boldsymbol{\beta}}, \sigma^2).$$

La log-vraisemblance, en omettant tout terme ou constante qui n'est pas fonction de σ^2, est

$$\ell(\widehat{\boldsymbol{\beta}}, \sigma^2) \propto -\frac{1}{2}\left\{n\ln\sigma^2 + \frac{1}{\sigma^2}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\}.$$

En différenciant chaque terme par rapport à σ^2 et en fixant le gradient à zéro, on obtient l'estimateur du maximum de vraisemblance

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{\mathsf{SS}_e}{n};$$

où SS_e est la somme des carrés des résidus. L'estimateur sans biais habituel de σ^2 calculé par le logiciel est $S^2 = \mathsf{SS}_e/(n - p - 1)$, où le dénominateur est la taille de l'échantillon n moins le nombre de paramètres de la moyenne $\boldsymbol{\beta}$, soit $p + 1$.

Proposition 3.1 (Invariance des estimateurs du maximum de vraisemblance)**.** *Si $g(\boldsymbol{\theta}) : \mathbb{R}^p \mapsto \mathbb{R}^k$ pour $k \leq p$ est une fonction des paramètres, alors $g(\widehat{\boldsymbol{\theta}})$ est l'estimateur du maximum de vraisemblance de cette fonction.*

La propriété d'invariance explique l'utilisation répandue de l'estimation du maximum de vraisemblance. Par exemple, après avoir estimé le paramètre λ, nous pouvons maintenant utiliser le modèle pour dériver d'autres quantités d'intérêt et obtenir les "meilleures" estimations gratuitement. Par exemple, nous pourrions calculer l'estimation du maximum de vraisemblance de la probabilité d'attendre plus d'une minute, $\Pr(T > 60) = \exp(-60/\widehat{\lambda}) = 0.126$. On peut utiliser la fonction de répartition `pexp` dans **R**,

```
# Note: la paramétrisation usuelle dans R pour la loi exponentielle
# est en terme d'intensité (réciproque du paramètre d'échelle)
pexp(q = 60, rate = 1/mean(attente), lower.tail = FALSE)
#> [1] 0.126
```

Un autre intérêt de la propriété d'invariance est la possibilité de calculer l'EMV dans la paramétrisation la plus simple, ce qui est pratique si le support est contraint. Si g est une fonction bijective de $\boldsymbol{\theta}$, par exemple si $\theta > 0$, maximiser le modèle paramétré en terme de $g(\theta) = \ln \theta$ ou de $g(\theta) = \ln(\theta) - \ln(1-\theta) \in \mathbb{R}$ si $0 \leq \theta \leq 1$, élimine les contraintes pour l'optimisation numérique.

Définition 3.3 (Score et information)**.** Soit $\ell(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, la fonction de log-vraisemblance. Le gradient (ou vecteur de dérivée première) de la log-vraisemblance $U(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ est appelé fonction de **score**.

L'**information observée** est la hessienne (matrice de dérivée deuxième) du négatif de la log-vraisemblance

$$j(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

En pratique, on évalue cette fonction à l'estimation du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}$, d'où le terme information observée pour désigner plutôt $j(\widehat{\boldsymbol{\theta}})$. Sous des conditions de régularité, l'**information de Fisher** est

$$i(\boldsymbol{\theta}) = \mathsf{E}\left\{U(\boldsymbol{\theta}; \boldsymbol{Y}) U(\boldsymbol{\theta}; \boldsymbol{Y})^\top\right\} = \mathsf{E}\left\{j(\boldsymbol{\theta}; \boldsymbol{Y})\right\}$$

La différence est qu'on prend l'espérance de chaque fonction des observations à l'intérieur des entrées de la matrice. Quand elle évaluée au point $\widehat{\boldsymbol{\theta}}$, l'information de Fisher mesure la variance du score, ou la courbure de ce dernier. La matrice de Fisher et la matrice d'information sont toutes deux symmétriques.

Exemple 3.6 (Information pour le modèle exponentiel)**.** L'information de Fisher et observée pour un échantillon aléatoire simple du modèle exponentiel, $Y_1, \ldots, Y_n$,, paramétré en terme d'échelle λ, est

$$j(\lambda; \boldsymbol{y}) = -\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = \frac{n}{\lambda^2} + \frac{2}{n\lambda^3}\sum_{i=1}^{n} y_i$$
$$i(\lambda) = \frac{n}{\lambda^2} + \frac{2}{n\lambda^3}\sum_{i=1}^{n} \mathsf{E}(Y_i) = \frac{n}{\lambda^2}$$

puisque $\mathsf{E}(Y_i) = \lambda$ et que l'espérance est un opérateur linéaire. On trouve que $i(\widehat{\lambda}) = j(\widehat{\lambda}) = n/\overline{y}^2$, mais cette égalité ne tient qu'à l'EMV.

Le modèle exponentiel peut s'avérer restrictif pour adéquatement capturer nos données, c'est pourquoi nous considérons une loi de Weibull comme généralisation.

Définition 3.4 (Loi de Weibull)**.** La fonction de répartition d'une variable aléatoire de loi **Weibull**, de paramètres d'échelle $\lambda > 0$ et de forme $\alpha > 0$ est

$$F(x; \lambda, \alpha) = 1 - \exp\left\{-(x/\lambda)^{\alpha}\right\}, \qquad x \geq 0, \lambda > 0, \alpha > 0,$$

alors que sa densité est

$$f(x; \lambda, \alpha) = \frac{\alpha}{\lambda^{\alpha}} x^{\alpha-1} \exp\left\{-(x/\lambda)^{\alpha}\right\}, \qquad x \geq 0, \lambda > 0, \alpha > 0.$$

La fonction quantile, qui est l'inverse de la fonction de répartition, est $Q(p) = \lambda\{-\ln(1-p)\}^{1/\alpha}$. La loi Weibull inclut la loi exponentielle comme cas spécial quand $\alpha = 1$. L'espérance de $Y \sim \mathsf{Weibull}(\lambda, \alpha)$ est $\mathsf{E}(Y) = \lambda\Gamma(1 + 1/\alpha)$.

Exemple 3.7 (Score et information d'une loi Weibull)**.** La log-vraisemblance d'un échantillon aléatoire simple de taille n dont la réalisation est dénotée $y_1, \ldots, y_n$, tirée d'une loi $\mathsf{Weibull}(\lambda, \alpha)$, est

$$\ell(\lambda, \alpha) = n\ln(\alpha) - n\alpha\ln(\lambda) + (\alpha - 1)\sum_{i=1}^{n} \ln y_i - \lambda^{-\alpha}\sum_{i=1}^{n} y_i^{\alpha}.$$

Le gradient de cette fonction est[3]

$$U(\lambda, \alpha) = \begin{pmatrix} \frac{\partial \ell(\lambda,\alpha)}{\partial \lambda} \\ \frac{\partial \ell(\lambda,\alpha)}{\partial \alpha} \end{pmatrix} = \begin{pmatrix} -\frac{n\alpha}{\lambda} + \alpha\lambda^{-\alpha-1}\sum_{i=1}^{n} y_i^{\alpha} \\ \frac{n}{\alpha} - n\ln(\lambda) + \sum_{i=1}^{n}\ln y_i - \sum_{i=1}^{n}\left(\frac{y_i}{\lambda}\right)^{\alpha} \times \ln\left(\frac{y_i}{\lambda}\right). \end{pmatrix}$$

[3]Par exemple, en utilisant une calculatrice symbolique.

et l'information observée est

$$\begin{aligned} j(\lambda, \alpha) &= -\begin{pmatrix} \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \lambda^2} & \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \lambda \partial \alpha} \\ \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \alpha \partial \lambda} & \frac{\partial^2 \ell(\lambda,\alpha)}{\partial \alpha^2} \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{-2}\left\{-n\alpha + \alpha(\alpha+1)\sum_{i=1}^n (y_i/\lambda)^2\right\} & \lambda^{-1}\sum_{i=1}^n [1-(y_i/\lambda)^\alpha\{1+\alpha\ln(y_i/\lambda)\}] \\ \lambda^{-1}\sum_{i=1}^n [1-(y_i/\lambda)^\alpha\{1+\alpha\ln(y_i/\lambda)\}] & n\alpha^{-2} + \sum_{i=1}^n (y_i/\lambda)^\alpha \{\ln(y_i/\theta)\}^2 \end{pmatrix} \end{aligned}$$

Proposition 3.2 (Optimisation basée sur le gradient)**.** *Pour obtenir l'estimateur du maximum de vraisemblance, nous trouverons généralement la valeur du vecteur $\boldsymbol{\theta}$ qui résout le vecteur de score, c'est-à-dire $U(\widehat{\boldsymbol{\theta}}) = \mathbf{0}_p$. Cela revient à résoudre simultanément un système de p équations en fixant à zéro la dérivée première par rapport à chaque élément de $\boldsymbol{\theta}$. Si $j(\widehat{\boldsymbol{\theta}})$ est une matrice définie positive (c'est-à-dire que toutes ses valeurs propres sont positives), alors le vecteur $\widehat{\boldsymbol{\theta}}$ maximise la fonction de log-vraisemblance et est l'estimateur du maximum de vraisemblance.*

Nous pouvons utiliser une variante de l'algorithme de Newton–Raphson si la vraisemblance est trois fois différentiable et si l'estimateur du maximum de vraisemblance ne se trouve pas sur la frontière de l'espace des paramètres. Si nous considérons une valeur initiale $\boldsymbol{\theta}^\dagger$, alors une expansion en série de Taylor du premier ordre de la vraisemblance du score dans un voisinage $\boldsymbol{\theta}^\dagger$ de l'EMV $\widehat{\boldsymbol{\theta}}$ donne

$$\begin{aligned} \mathbf{0}_p = U(\widehat{\boldsymbol{\theta}}) &\stackrel{\cdot}{\simeq} \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\dagger} + \left.\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\dagger} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\dagger) \\ &= U(\boldsymbol{\theta}^\dagger) - j(\boldsymbol{\theta}^\dagger)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\dagger). \end{aligned}$$

En réarrangeant cette expression pour isoler $\widehat{\boldsymbol{\theta}}$ (pourvu que la matrice $p \times p$ d'information observée $j(\widehat{\boldsymbol{\theta}})$ soit inversible à l'EMV), on obtient

$$\widehat{\boldsymbol{\theta}} \stackrel{\cdot}{\simeq} \boldsymbol{\theta}^\dagger + j^{-1}(\boldsymbol{\theta}^\dagger) U(\boldsymbol{\theta}^\dagger).$$

Cela suggère l'utilisation d'une procédure itérative: à partir d'une valeur de départ $\boldsymbol{\theta}^\dagger$ dans le voisinage du mode, on applique le schéma de mise à jour jusqu'à ce que le gradient soit approximativement nul. Si la valeur est éloignée du mode, l'algorithme peut diverger. Pour éviter cela, nous pouvons multiplier le terme $j^{-1}(\boldsymbol{\theta}^\dagger)U(\boldsymbol{\theta}^\dagger)$ par un facteur d'amortissement $c < 1$. Une variante de l'algorithme, appelée score de Fisher, utilise l'information de Fisher $i(\boldsymbol{\theta})$ au lieu de l'information observée, $j(\boldsymbol{\theta})$, pour des raisons de stabilité numérique et pour éviter les situations où cette dernière n'est pas définie positive. Il s'agit de la routine d'optimisation utilisée dans la fonction `glm` *de* ***R****.*

Exemple 3.8 (Estimateurs du maximum de vraisemblance d'un échantillon Weibull)**.** Nous nous tournons vers l'optimisation numérique pour obtenir l'estimation du maximum de vraisemblance de la loi de Weibull, en l'absence formule explicite pour les

EMV. À cette fin, il faut écrire une fonction qui encodent la log-vraisemblance, ici la somme des contributions de la log-densité. La fonction `nll_weibull` ci-dessous prend comme premier argument le vecteur de paramètres, `pars`, et renvoie la valeur négative de la log-vraisemblance que nous souhaitons minimiser[4]. Nous codons également le gradient, bien que nous puissions recourir à la différenciation numérique. Nous utilisons ensuite `optim`, la routine d'optimisation par défaut de **R**, pour minimiser `nll_weibull`. La fonction renvoie une liste contenant un code de convergence (`0` indiquant la convergence), les EMV dans `par`, la log-vraisemblance $\ell(\widehat{\boldsymbol{\theta}})$ et la hessienne, qui est la matrice d'information observée évaluée à $\widehat{\boldsymbol{\theta}}$. La surface de log-vraisemblance, pour les paires de vecteurs d'échelle et de forme $\boldsymbol{\theta} = (\lambda, \alpha)$, est représentée dans la Figure 3.3. Nous pouvons voir que l'algorithme a convergé vers le maximum de vraisemblance et vérifier que le score satisfait $U(\widehat{\boldsymbol{\theta}}) = 0$ à la valeur optimale retournée.

```
# Charger les données
data(attente, package = "hecstatmod")

# Négatif de la log vraisemblance pour un échantillon Weibull
nll_weibull <- function(pars, y) {
  # Gérer le cas de paramètres négatifs (impossible)
  if (isTRUE(any(pars <= 0))) {
    return(1e10) # retourner une valeur large finie (pour éviter les messages d'avertissemen
  }
  - sum(dweibull(
    x = y,
    scale = pars[1],
    shape = pars[2],
    log = TRUE
  ))
}
# Gradient du négatif de la fonction de log vraisemblance Weibull
gr_nll_weibull <- function(pars, y) {
  scale <- pars[1]
  shape <- pars[2]
  n <- length(y)
  grad_ll <- c(
    scale = -n * shape / scale + shape * scale^(-shape - 1) * sum(y^shape),
    shape = n / shape - n * log(scale) + sum(log(y)) -
```

[4] La plupart des algorithmes d'optimisation minimisent les fonctions par rapport à leurs arguments, nous minimisons donc la log-vraisemblance négative, ce qui équivaut à maximiser la log-vraisemblance

```
      sum(log(y / scale) * (y / scale)^shape)
  )
  return(-grad_ll)
}

# Utiliser les EMV du modèle exponentiel pour l'initialisation
valinit <- c(mean(attente), 1)
# Vérifier préalablement que le gradient est correct!
# La commande retourne TRUE si la dérivée numérique égale sa version analytique à tolérance
isTRUE(all.equal(
  numDeriv::grad(nll_weibull, x = valinit, y = attente),
  gr_nll_weibull(pars = valinit, y = attente),
  check.attributes = FALSE
))
#> [1] TRUE
# Optimisation numérique avec optim
opt_weibull <- optim(
  par = valinit,
  # valeurs initiales
  fn = nll_weibull,
  # passer la fonction à optimiser, son premier argument doit être le vecteur de paramètres
  gr = gr_nll_weibull,
  # gradient (optionnel)
  method = "BFGS",
  # algorithme BFGS est basé sur le gradient, une alternative robuste est"Nelder"
  y = attente,
  # vecteur d'observations passées en argument additionnel à "fn"
  hessian = TRUE # retourner la matrice de dérivée secondes évaluée aux EMV
)
# Alternative avec un Newton
# nlm(f = nll_weibull, p = valinit, hessian = TRUE, y = attente)
# Estimations du maximum de vraisemblance
(mle_weibull <- opt_weibull$par)
#> [1] 32.6  2.6
# Vérifier la convergence numérique à l'aide du gradient
gr_nll_weibull(mle_weibull, y = attente)
#>     scale     shape
#> 0.0000142 0.0001136
# Vérifier que la hessienne est positive définite
```

```
# Toutes les valeurs propres sont positives
# Si oui, on a trouvé un maximum et la matrice est invertible
isTRUE(all(eigen(opt_weibull$hessian)$values > 0))
#> [1] TRUE
```

3.2 Loi d'échantillonnage

La **loi d'échantillonnage** d'un estimateur $\widehat{\boldsymbol{\theta}}$ est la loi de probabilité induite par les données aléatoires sous-jacentes.

Supposons que nous disposons d'un échantillon aléatoire simple, de sorte que la log-vraisemblance est constitutée d'une somme de n termes et que l'information s'accumule linéairement avec la taille de l'échantillon. Nous dénotons la vraie valeur du vecteur de paramètres inconnu $\boldsymbol{\theta}_0$. Sous des conditions de régularité appropriées, cf. section 4.4.2 de Davison (2003), pour un échantillon de grande taille n, nous pouvons effectuer une série de Taylor du score et appliquer le théorème de la limite centrale à la moyenne résultante puisque $U(\boldsymbol{\theta})$ et $i(\boldsymbol{\theta})$ sont la somme de n variables aléatoires indépendantes, et que $\mathsf{E}\{U(\boldsymbol{\theta})\} = \mathbf{0}_p$, et $\mathsf{Var}\{U(\boldsymbol{\theta})\} = i(\boldsymbol{\theta})$, l'application du théorème de la limite centrale et de la loi des grands nombres donne

$$i(\boldsymbol{\theta}_0)^{-1/2}U(\boldsymbol{\theta}_0) \dot{\sim} \mathsf{normale}_p(\mathbf{0}, \mathbf{I}_p).$$

On peut utiliser ce résultat pour obtenir une approximation à la loi d'échantillonnage des estimateurs du maximum de vraisemblance de $\widehat{\boldsymbol{\theta}}$,

$$\widehat{\boldsymbol{\theta}} \dot{\sim} \mathsf{normale}_p\{\boldsymbol{\theta}_0, i^{-1}(\boldsymbol{\theta})\}$$

ou la matrice de covariance est l'inverse de l'information de Fisher. En pratique, puisque la valeur des paramètres $\boldsymbol{\theta}_0$ est inconnue, on remplace la covariance soit par $i^{-1}(\widehat{\boldsymbol{\theta}})$ ou par l'inverse de l'information observée, $j^{-1}(\widehat{\boldsymbol{\theta}})$. Cela est justifié par le fait que les deux matrices d'informations $i(\widehat{\boldsymbol{\theta}})$ et $j(\widehat{\boldsymbol{\theta}})$ convergent vers $i(\boldsymbol{\theta})$ quand $n \to \infty$.

Au fur et à mesure que la taille de l'échantillon augmente, l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}$ devient centré autour de la valeur $\boldsymbol{\theta}_0$ qui minimise l'écart entre le modèle et le véritable processus de génération des données. Dans les grands échantillons, la loi d'échantillonnage de l'estimateur du maximum de vraisemblance est approximativement quadratique.

Exemple 3.9 (Matrice de covariance et erreurs-type pour le modèle de Weibull)**.** Nous utilisons la sortie de notre procédure d'optimisation pour obtenir la matrice d'information observée et les erreurs-type pour les paramètres du modèle de Weibull. Ces dernières sont simplement la racine carrée des entrées diagonales de l'information observée évaluée aux EMV, $[\mathrm{diag}\{j^{-1}(\widehat{\boldsymbol{\theta}})\}]^{1/2}$.

```
# La hessienne du négatif de la log vraisemblance, évaluée aux EMV
# est la matrice d'information observée
obsinfo_weibull <- opt_weibull$hessian
vmat_weibull <- solve(obsinfo_weibull)
# Erreurs-type
se_weibull <- sqrt(diag(vmat_weibull))
```

Une fois que l'on a les estimations du maximum de vraisemblance et les erreurs-type, on peut dériver des intervalles de confiance ponctuels de Wald pour les paramètres de $\boldsymbol{\theta}$. Si la quantité d'intérêt est une transformation des paramètres du modèle, on peut utiliser le résultat suivant pour procéder.

Proposition 3.3 (Normalité asymptotique et transformations)**.** *Le résultat de normalité asymptotique peut être utilisé pour dériver les erreurs standard pour d'autres quantités d'intérêt. Si $\phi = g(\boldsymbol{\theta})$ est une fonction différentiable de $\boldsymbol{\theta}$ dont le gradient est non-nul lorsque évalué à $\widehat{\boldsymbol{\theta}}$, alors $\widehat{\phi} \dot{\sim} \mathsf{normale}(\phi_0, \mathrm{V}_\phi)$, with $\mathrm{V}_\phi = \nabla\phi^\top \mathbf{V}_{\boldsymbol{\theta}} \nabla\phi$, où $\nabla\phi = [\partial\phi/\partial\theta_1, \ldots, \partial\phi/\partial\theta_p]^\top$. La matrice de covariance et le gradient sont évalués aux estimations du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}$. Ce résultat se généralise aux fonctions vectorielles $\boldsymbol{\phi} \in \mathbb{R}^k$ pour $k \leq p$, où $\nabla\phi$ est la jacobienne de la transformation.*

Exemple 3.10 (Probabilité d'attente pour un modèle exponentiel.)**.** Considérons les données sur le temps d'attente dans le métro et la probabilité d'attendre plus d'une minute, $\phi = g(\lambda) = \exp(-60/\lambda)$. L'estimation du maximum de vraisemblance est, par invariance, 0.126 et le gradient de g par rapport au paramètre d'échelle est $\nabla\phi = \partial\phi/\partial\lambda = 60\exp(-60/\lambda)/\lambda^2$.

```
# Exemple de dérivation des erreurs-type pour une
# transformation des paramètres
# Ici, on calcule Pr(Y>60) selon le modèle exponentiel
lambda_hat <- mean(attente)
# Définir la fonction d'intérêt
```

```
phi_hat <- exp(-60 / lambda_hat)
# jacobien de la transformation
dphi <- function(lambda) {
  60 * exp(-60 / lambda) / (lambda^2)
}
# variance du paramètre exponentiel
V_lambda <- lambda_hat^2 / length(attente)
# variance de Pr(Y>60) via la méthode delta
V_phi <- dphi(lambda_hat)^2 * V_lambda
# extraire et imprimer les erreurs-type
(se_phi <- sqrt(V_phi))
#> [1] 0.0331
```

3.3 Tests dérivés de la vraisemblance

Nous considérons une hypothèse nulle $\mathscr{H}_0$ qui impose des restrictions sur les valeurs possibles de $\boldsymbol{\theta}$, par rapport à une alternative sans contrainte $\mathscr{H}_1$. Nous avons besoin de deux modèles **emboîtés** : un modèle *complet* et un modèle *réduit*, pour lequel l'espace des paramèteres est un sous-ensemble du modèle complet suite à l'imposition des q restrictions. Par exemple, la loi exponentielle est un cas particulier de la loi de Weibull si $\alpha = 1$.

L'hypothèse nulle $\mathscr{H}_0$ testée est "le modèle réduit est une **simplification adéquate** du modèle complet". Soit $\widehat{\boldsymbol{\theta}}_0$ les EMV contraints pour le modèle sous l'hypothèse nulle, et $\widehat{\boldsymbol{\theta}}$ les EMV du modèle complet. La vraisemblance fournit trois classes principales de statistiques pour tester cette hypothèse, soit

- les statistiques des tests du rapport de vraisemblance, notées R, qui mesurent la différence de log vraisemblance (distance verticale) entre $\ell(\widehat{\boldsymbol{\theta}})$ et $\ell(\widehat{\boldsymbol{\theta}}_0)$.
- les statistiques des tests de Wald, notées W, qui considèrent la distance horizontale normalisée entre $\widehat{\boldsymbol{\theta}}$ et $\widehat{\boldsymbol{\theta}}_0$.
- les statistiques des tests de score de Rao, notées S, qui examinent le gradient repondéré de ℓ, évaluée *uniquement* à $\widehat{\boldsymbol{\theta}}_0$.

Les trois principales classes de statistiques permettant de tester une hypothèse nulle

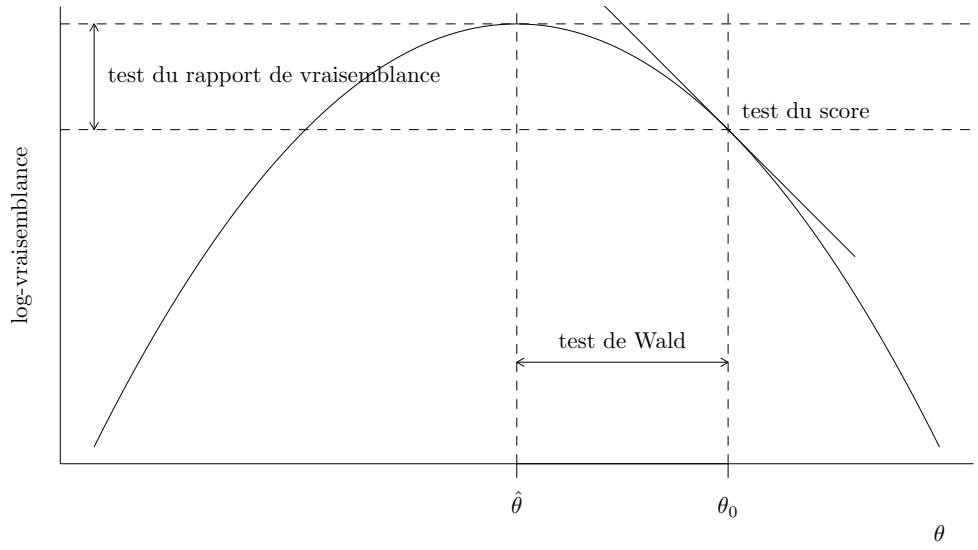

Figure 3.2: Fonction de log vraisemblance et illustrations des éléments des statistique du score, de Wald et du rapport de vraisemblance.

simple $\mathscr{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ par rapport à l'hypothèse alternative $\mathscr{H}_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ sont

$$\begin{aligned} W(\boldsymbol{\theta}_0) &= (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top j(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), && \text{(Wald)} \\ R(\boldsymbol{\theta}_0) &= 2\left\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\right\}, && \text{(rapport de vraisemblance)} \\ S(\boldsymbol{\theta}_0) &= U^\top(\boldsymbol{\theta}_0) i^{-1}(\boldsymbol{\theta}_0) U(\boldsymbol{\theta}_0), && \text{(score)} \end{aligned}$$

où $\boldsymbol{\theta}_0$ est la valeur nulle postulée du paramètre avec q restrictions. Si $q \neq p$, alors on remplace $\boldsymbol{\theta}_0$ par l'estimation contrainte $\widehat{\boldsymbol{\theta}}_0$.

Asymptotiquement, toutes les statistiques de test sont équivalentes (dans le sens où elles conduisent aux mêmes conclusions sur $\mathscr{H}_0$), mais elles ne sont pas identiques. Sous $\mathscr{H}_0$, les trois statistiques de test suivent une loi asymptotique χ^2_q, où les degrés de liberté q indiquent le nombre de restrictions.

Si θ est un scalaire (cas $q = 1$), des versions directionnelles de ces statistiques existent,

$$\begin{aligned} w(\theta_0) &= (\widehat{\theta} - \theta_0)/\mathsf{se}(\widehat{\theta}) && \text{(Wald)} \\ r(\theta_0) &= \operatorname{sign}(\widehat{\theta} - \theta)\left[2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}\right]^{1/2} && \text{(racine directionnelle de vraisemblance)} \\ s(\theta_0) &= i^{-1/2}(\theta_0) U(\theta_0) && \text{(score)} \end{aligned}$$

Sous cette forme, si l'hypothèse nulle $\mathscr{H}_0 : \theta = \theta_0$ est vraie, alors $w(\theta_0) \dot{\sim} \text{normale}(0,1)$, etc.

La statistique du test du rapport de vraisemblance est normalement la plus puissante des trois tests (et donc préférable selon ce critère); la statistique est aussi invariante aux reparamétrages. La statistique de score S, moins utilisée, nécessite le calcul du score et de l'information de Fisher, mais n'est évaluée que sous $\mathscr{H}_0$ (car par définition $U(\widehat{\theta}) = 0$), elle peut donc être utile dans les problèmes où les calculs de l'estimateur du maximum de vraisemblance sous l'alternative sont coûteux ou impossibles. Le test de Wald est le plus facile à dériver, mais son taux de couverture empirique peut laisser à désirer si la loi d'échantillonnage de $\widehat{\boldsymbol{\theta}}$ est fortement asymétrique.

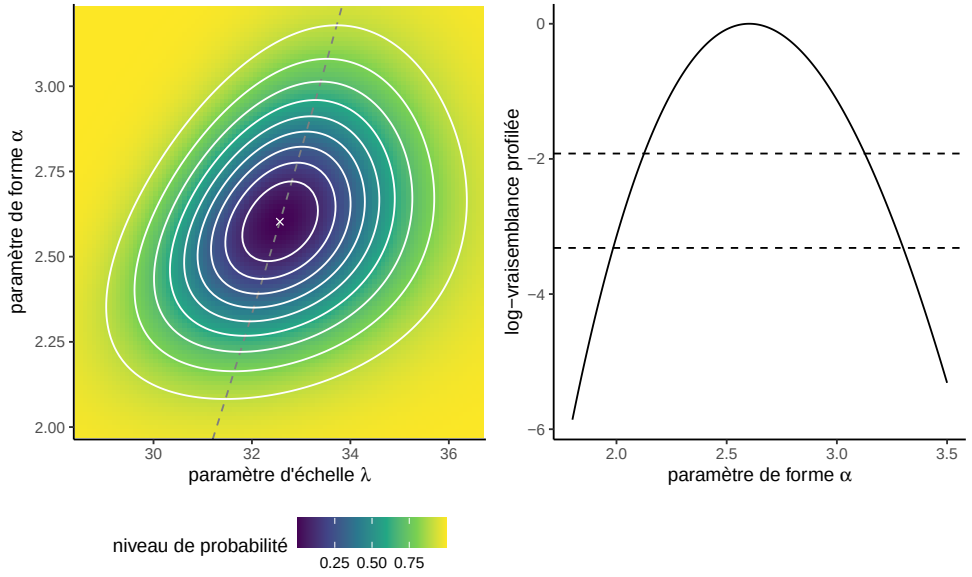

Figure 3.3: Log-vraisemblance profilée pour α, représentée par un trait gris traitillé (gauche) et par une coupe transversale (droite). Le panneau de gauche montre la surface de log-vraisemblance pour le modèle de Weibull avec des régions de confiance de 10%, 20%, ..., 90% du rapport de vraisemblance (courbes de contour blanches). Les valeurs de log vraisemblance les plus élevées sont indiquées par des couleurs plus foncées, et la valeur des estimations du maximum de vraisemblance par une croix. La vraisemblance profilée du panneau de droite a été décalée verticalement pour que sa valeur maximale soit zéro; les lignes horizontales traitillées indiquent les valeurs pour les intervalles de confiance à 95% et 99%.

La statistique de Wald W est la plus courante. Les intervalles de confiance bilatéraux de

niveau $(1-\alpha)$ de Wald pour les paramètres de $\boldsymbol{\theta}$, où pour θ_j $(j=1,\ldots,p)$,

$$\widehat{\theta}_j \pm \mathfrak{z}_{1-\alpha/2}\mathrm{se}(\widehat{\theta}_j),$$

avec $\mathfrak{z}_{1-\alpha/2}$ le quantile $1-\alpha/2$ d'une loi normale standard. Pour un intervalle à 95%, le 0.975 quantile vaut $\mathfrak{z}_{0.975}=1.96$. Les intervalles de confiance de Wald bilatéraux sont, par construction, symmétriques. Parfois, cela donne des valeurs impossibles (par exemple, une variance négative).

Exemple 3.11 (Test de Wald pour comparer les modèles Weibull et exponentiel)**.** Nous pouvons tester si la loi exponentielle est une simplification adéquate de la loi de Weibull en imposant la restriction $\mathscr{H}_0 : \alpha = 1$. Nous comparons les statistiques de Wald W à un χ^2_1. Puisque α est un paramètre de la loi Weibull, nous avons les erreurs-type gratuitement.

```
# Calculer la statistique de Wald
wald_exp <- (mle_weibull[2] - 1)/se_weibull[2]
# Calculer la valeur-p
pchisq(wald_exp^2, df = 1, lower.tail = FALSE)
#> [1] 3.61e-10
# valeur-p inférieure à 5%, rejet de l'hypothèse nulle
# Intervalles de confiance de niveau 95%
mle_weibull[2] + qnorm(c(0.025, 0.975))*se_weibull[2]
#> [1] 2.1 3.1
# La valeur 1 n'appartient pas à l'intervalle, rejeter H0
```

Nous rejetons l'hypothèse nulle, ce qui signifie que le sous-modèle exponentiel n'est pas une simplification adéquate du modèle de Weibull $(\alpha \neq 1)$.

Nous pouvons également vérifier l'ajustement des deux modèles à l'aide d'un diagramme quantile-quantile (cf. Définition 1.14). Il ressort de Figure 3.4 que le modèle exponentiel surestime les temps d'attente les plus importants, dont la dispersion dans l'échantillon est inférieure à celle impliquée par le modèle. En revanche, la ligne droite presque parfaite pour le modèle de Weibull dans le panneau de droite de Figure 3.4 suggère que l'ajustement du modèle est adéquat.

Remarque 3.1 (Absence d'invariance des intervalles de confiance de Wald). Puisque les erreurs-types de paramètres dépendent de la paramétrisation, les intervalles de confiance de Wald ne sont pas invariants à ces transformations. Par exemple, si on veut des intervalles de confiance pour une fonction $g(\boldsymbol{\theta})$ qui n'est pas linéaire, alors en général. $\mathsf{IC}_W\{g(\theta)\} \neq g\{\mathsf{IC}_W(\theta)\}$.

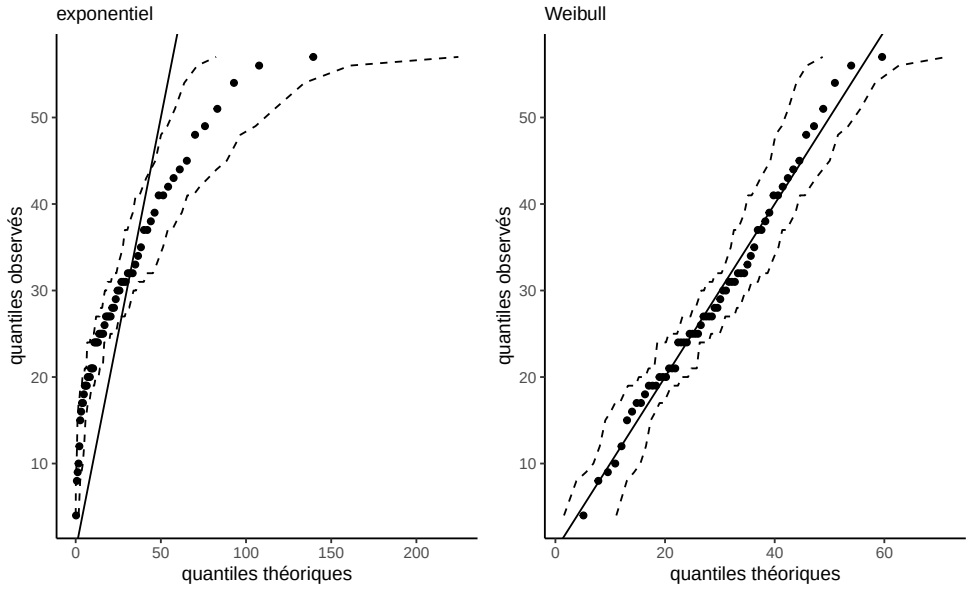

Figure 3.4: Diagrammes quantile-quantile des modèles exponentiel (gauche) et Weibull (droite) avec intervalles de confiance ponctuels à 95% obtenus par autoamorçage.

Par exemple, considérons le modèle exponentiel. Nous pouvons inverser la statistique du test de Wald pour obtenir un intervalle de confiance symétrique à 95% pour $\phi = g(\lambda) = \exp(-60/\lambda)$, $ [0.061,$ $0.191]$. Si nous devions naïvement transformer l'intervalle de confiance pour λ en un pour ϕ en appliquant la fonction $g(\cdot)$ à chaque borne, nous obtiendrions plutôt $[0.063,\ 0.19]$, Bien que la différence soit minime ici, cela met en évidence l'invariance. L'approximation gaussienne qui sous-tend le test de Wald est fiable si la loi d'échantillonnage de la vraisemblance est presque quadratique, ce qui se produit lorsque la fonction de vraisemblance est à peu près symétrique de part et d'autre de l'estimateur du maximum de vraisemblance.

Le test du rapport de vraisemblance est invariant par rapport aux reparamétrages préservant les intérêts, de sorte que la statistique de test pour $\mathscr{H}_0 : \phi = \phi_0$ et $\mathscr{H}_0 : \lambda = -60/\ln(\phi_0)$ est la même. Les intervalles de confiance de Wald peuvent être comparées à celles (meilleures) obtenues à l'aide du test du rapport de vraisemblance. Ces dernières sont obtenues par une recherche numérique des limites de

$$\left\{\theta : 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})\} \leq \chi^2_p(1-\alpha)\right\},$$

où $\chi^2_p(1-\alpha)$ est le quantile de niveau $(1-\alpha)$ de la loi χ^2_p. De tels intervalles, pour $\alpha = 0.1, \ldots, 0.9$, sont tracés sur la Figure 3.3 (courbes de contour). Si $\boldsymbol{\theta}$ est un p-vecteur $(p > 1)$,

alors les intervalles de confiance pour θ_i sont dérivés à partir de la vraisemblance profilée. Les intervalles de confiance basés sur la statistique du rapport de vraisemblance sont **invariants aux reparamétrages**, donc $\mathsf{IC}_R\{g(\theta)\} = g\{\mathsf{IC}_R(\theta)\}$. Comme la vraisemblance est nulle si la valeur d'un paramètre se situe en dehors de l'espace des paramètres $\boldsymbol{\Theta}$, les intervalles n'incluent que les valeurs plausibles de θ. En général, les intervalles sont asymétriques et présentent de meilleures taux de couverture.

```
# Log vraisemblance exponentielle
ll_exp <- function(lambda) {
  sum(dexp(attente, rate = 1 / lambda, log = TRUE))
}
# EMV du paramètre d'échelle
lambda_hat <- mean(attente)
# Recherche des zéros de la fonction pour obtenir
# les limites des intervalles de confiance
lrt_lb <- uniroot(
  # borne inférieure, en utilisant l'EMV
  f = function(r) {
    2 * (ll_exp(lambda_hat) - ll_exp(r)) - qchisq(0.95, 1)
  },
  interval = c(0.5 * min(attente), lambda_hat)
)$root
lrt_ub <- uniroot(
  # borne supérieure
  f = function(r) {
    2 * (ll_exp(lambda_hat) - ll_exp(r)) - qchisq(0.95, 1)
  },
  interval = c(lambda_hat, 2 * max(attente))
)$root
```

L'intervalle de confiance à 95% de la statistique du rapport de vraisemblance pour λ peut être trouvé en utilisant un algorithme de recherche linéaire: l'intervalle de confiance à 95% pour λ est $\mathsf{IC}_R(\lambda)[22.784, 37.515]$. Par invariance, l'intervalle de confiance à 95% pour ϕ est $\mathsf{IC}_R(\phi) = [0.072, 0.202] = g\{\mathsf{IC}_R(\lambda)\}$.

3.4 Vraisemblance profilée

Parfois, nous pouvons vouloir effectuer des tests d'hypothèse ou dériver des intervalles de confiance pour un sous-ensemble spécifique des paramètres du modèle, ou une transformation de ces derniers. Dans ce cas, l'hypothèse nulle ne restreint qu'une partie de l'espace et les autres paramètres, dits de nuisance, ne sont pas spécifiés — la question est alors de savoir quelles valeurs utiliser pour la comparaison avec le modèle complet. Il s'avère que les valeurs qui maximisent la log-vraisemblance contrainte sont celles que l'on doit utiliser pour le test, et la fonction particulière dans laquelle ces paramètres de nuisance sont intégrés est appelée vraisemblance profilée.

Définition 3.5 (Log-vraisemblance profilée)**.** Soit un modèle paramétrique avec log-vraisemblance $\ell(\boldsymbol{\theta})$, dont le vecteur de paramètres de dimension p $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\varphi})$ peut être séparé en un sous-vecteur de longueur q contenant les paramètres d'intérêts, disons $\boldsymbol{\psi}$ et un sous-vecteur de longueur $(p-q)$ contenant les paramètres de nuisance $\boldsymbol{\varphi}$.

La log-vraisemblance profilée ℓ_{p} est une fonction de $\boldsymbol{\psi}$ qui est obtenue en maximisant la log-vraisemblance ponctuellement à chaque valeur fixe $\boldsymbol{\psi}_0$ sur le vecteur de nuisance $\boldsymbol{\varphi}_{\psi_0}$,

$$\ell_{\mathsf{p}}(\boldsymbol{\psi}) = \max_{\boldsymbol{\varphi}} \ell(\boldsymbol{\psi}, \boldsymbol{\varphi}) = \ell(\boldsymbol{\psi}, \widehat{\boldsymbol{\varphi}}_{\boldsymbol{\psi}}).$$

Exemple 3.12 (Log-vraisemblance profilée pour le paramètre de forme d'une loi Weibull)**.** Considérons le paramètre de forme $\psi \equiv \alpha$ comme paramètre d'intérêt, et le paramètre d'échelle $\varphi \equiv \lambda$ comme paramètre de nuisance. En utilisant le gradient dérivé dans l'Exemple 3.7, nous constatons que la valeur de l'échelle qui maximise la log-vraisemblance pour un α donné est

$$\widehat{\lambda}_\alpha = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^\alpha\right)^{1/\alpha}.$$

Si on substitue cette valeur dans la log-vraisemblance, on obtient une fonction de α uniquement, ce qui réduit également le problème d'optimisation pour les EMV d'une loi Weibull à une recherche linéaire le long de $\ell_{\mathsf{p}}(\alpha)$. Le panneau de gauche de Figure 3.3 montre la crête le long de la direction de α correspondant à la surface de log-vraisemblance. Si l'on considère ces courbes de niveau comme celles d'une carte topographique, la log-vraisemblance profilée correspond dans ce cas à une marche le long de la crête des deux montagnes dans la direction ψ, le panneau de droite montrant le gain/la perte d'altitude. Le profil d'élévation correspondant à droite de Figure 3.3 avec les points de coupure pour les intervalles de confiance basés sur le rapport de vraisemblance<. Nous devrions obtenir numériquement, à l'aide d'un algorithme de recherche lin/aire, les limites de l'intervalle

de confiance de part et d'autre de $\widehat{\alpha}$, mais il est clair que $\alpha = 1$ n'est pas dans l'intervalle de 99%.

```
# EMV conditionnels de lambda pour alpha donné
lambda_alpha <- function(alpha, y = attente) {
  (mean(y^alpha))^(1 / alpha)
}
# Log vraisemblance profilée pour alpha
prof_alpha_weibull <- function(par, y = attente) {
  sapply(par, function(a) {
    nll_weibull(pars = c(lambda_alpha(a), a), y = y)
  })
}
```

Exemple 3.13 (Log-vraisemblance profilée pour l'espérance d'une loi Weibull)**.** Nous pouvons également utiliser l'optimisation numérique pour calculer la log-vraisemblance profilée d'une fonction des paramètres. Supposons que nous soyons intéressés par le temps moyen d'attente théorique. Selon le modèle Weibull, cette valeur est $\mu = \mathsf{E}(Y) = \lambda\Gamma(1 + 1/\alpha)$. À cet effet, nous reparamétrons le modèle en termes de (μ, α), où $\lambda = \mu/\Gamma(1 + 1/\alpha)$. Nous créons ensuite une fonction qui optimise la log-vraisemblance pour une valeur fixe de μ, puis renvoie $\widehat{\alpha}_\mu$, μ et $\ell_{\mathrm{p}}(\mu)$.

Pour obtenir les intervalles de confiance d'un paramètre scalaire, il existe une astuce qui permet de s'en tirer avec une évaluation sommaire, pour autant que la log-vraisemblance profilée soit relativement lisse. Nous calculons la racine directionnelle du rapport de vraisemblance, $r(\psi) = \mathrm{sign}(\psi - \widehat{\psi})\{2\ell_{\mathrm{p}}(\widehat{\psi}) - 2\ell_{\mathrm{p}}(\psi)\}^{1/2}$ sur une grille fine de valeurs de ψ, puis nous ajustons une spline de lissage, une régression avec variable réponse $y = \psi$ et variable explicative $x = r(\psi)$. Nous prédisons ensuite la courbe aux quantiles normaux $\mathfrak{z}_{\alpha/2}$ et $\mathfrak{z}_{1-\alpha/2}$, et renvoyons ces valeurs sous forme d'intervalle de confiance. La Figure 3.5 montre comment ces valeurs correspondent aux points de coupure sur l'échelle du logarithme du rapport de vraisemblance, où la ligne verticale est donnée par $-\mathfrak{c}(1-\alpha)/2$ où $\mathfrak{c}$ représente le quantile d'une variable aléatoire χ^2_1.

```
# Compute the MLE for the expected value via plug-in
mu_hat <- mle_weibull[1]*gamma(1+1/mle_weibull[2])
# Create a profile function
prof_weibull_mu <- function(mu){
  # For given value of mu
```

```
  alpha_mu <- function(mu){
  # Find the profile by optimizing (line search) for fixed mu and the best alpha
     opt <- optimize(f = function(alpha, mu){
     # minimize the negative log likelihood
      nll_weibull(c(mu/gamma(1+1/alpha), alpha), y = attente)},
   mu = mu,
  interval = c(0.1,10) #search region
  )
  # Return the value of the negative log likelihood and alpha_mu
  return(c(nll = opt$objective, alpha = opt$minimum))
  }
  # Create a data frame with mu and the other parameters
  data.frame(mu = mu, t(sapply(mu, function(m){alpha_mu(m)})))
}
# Create a data frame with the profile
prof <- prof_weibull_mu(seq(22, 35, length.out = 101L))
# Compute signed likelihood root r
prof$r <- sign(prof$mu - mu_hat)*sqrt(2*(prof$nll - opt_weibull$value))

# Trick: fit a spline to obtain the predictions with mu as a function of r
# Then use this to predict the value at which we intersect the normal quantiles
fit.r <- stats::smooth.spline(x = cbind(prof$r, prof$mu), cv = FALSE)
pr <- predict(fit.r, qnorm(c(0.025, 0.975)))$y
# Plot the signed likelihood root - near linear indicates quadratic
g1 <- ggplot(data = prof,
     mapping = aes(x = mu, y = r)) +
  geom_abline(intercept = 0, slope = 1) +
  geom_line() +
  geom_hline(yintercept = qnorm(0.025, 0.975),
             linetype = "dashed") +
  labs(x = expression(paste("espérance ", mu)),
       y = "racine directionnelle de vraisemblance")
# Create a plot of the profile
g2 <- ggplot(data = prof,
       mapping = aes(x = mu, y = opt_weibull$value - nll)) +
  geom_line() +
  geom_hline(yintercept = -qchisq(c(0.95), df = 1)/2,
             linetype = "dashed") +
  geom_vline(linetype = "dotted",
```

```
  xintercept = pr) +
  labs(x = expression(paste("espérance ", mu)),
       y = "log vraisemblance profilée")

g1 + g2
```

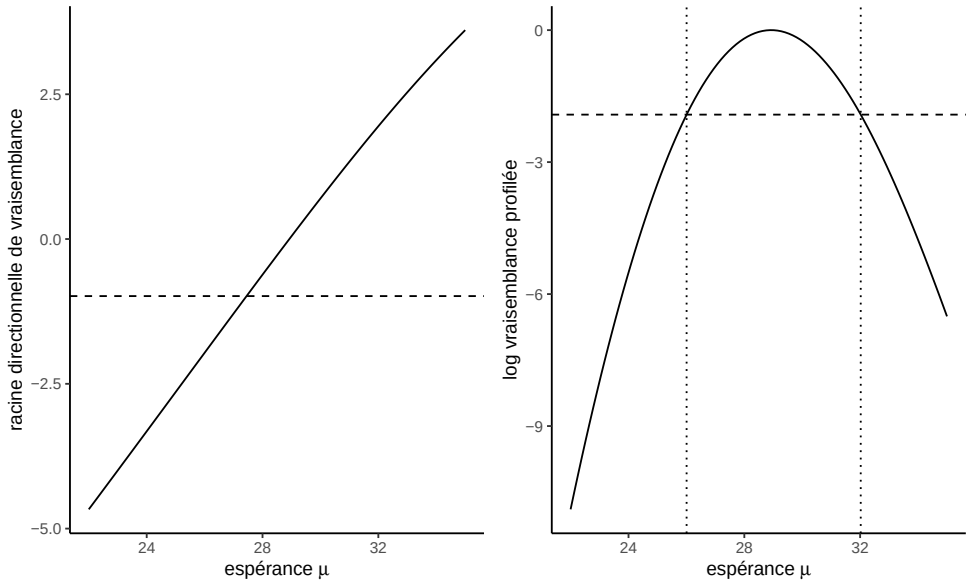

Figure 3.5: Racine directionnelle du rapport de vraisemblance (gauche) et log vraisemblance profilée (droite) en fonction de l'espérance μ pour un modèle Weibull.

L'estimateur du maximum de vraisemblance du profil se comporte comme une vraisemblance normale pour la plupart des quantités d'intérêt et nous pouvons dériver des statistiques de test et des intervalles de confiance de la manière habituelle. Un exemple célèbre de profil de vraisemblance est la fonction de risque proportionnel de Cox couvert dans le chapitre 7.

Exemple 3.14 (Transformation de Box–Cox)**.** Parfois, le postulat de normalité de l'erreur dans une régression linéaire ne tient pas. Si les données sont strictement positives, on peut envisager une transformation de Box-Cox,

$$y(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0. \end{cases}$$

Si on postule que $\boldsymbol{Y}(\lambda) \sim \mathsf{normale}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, alors la log-vraisemblance s'écrit

$$L(\lambda, \boldsymbol{\beta}, \sigma; \boldsymbol{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} J(\lambda, \boldsymbol{y}) \times \exp\left[-\frac{1}{2\sigma^2}\{\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta}\}^\top\{\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta}\}\right],$$

où J dénote le jacobien de la transformation de Box–Cox, $J(\lambda, \boldsymbol{y}) = \prod_{i=1}^n y_i^{\lambda-1}$. Pour chaque valeur de λ, l'estimateur du maximum de vraisemblance est le même que celle de la régression linéaire, mais où $\boldsymbol{y}$ est remplacée par $\boldsymbol{y}(\lambda)$, soit $\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}(\lambda)$ and $\widehat{\sigma}^2_\lambda = n^{-1}\{\boldsymbol{y}(\lambda) - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\}^\top\{\boldsymbol{y}(\lambda) - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\}$.

La log-vraisemblance profilée est donc

$$\ell_{\mathsf{p}}(\lambda) = -\frac{n}{2}\ln(2\pi\widehat{\sigma}^2_\lambda) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^n \ln(y_i)$$

L'estimateur du maximum de vraisemblance profilée est la valeur λ qui minimise la somme des carrés des résidus du modèle linéaire avec $\boldsymbol{y}(\lambda)$ comme réponse.

La transformation de Box–Cox n'est pas une solution miracle et doit être réservée aux cas où la transformation réduit l'hétéroscédasticité (variance inégale) ou crée une relation linéaire entre les explications et la réponse. La théorie fournit une explication convaincante des données avec, par exemple, la fonction de production Cobb-Douglas utilisée en économie qui peut être linéarisée par une transformation logarithmique. Plutôt que de choisir une transformation *ad hoc*, on pourrait choisir une transformation logarithmique si la valeur 0$ est incluse dans l'intervalle de confiance à 95%, car cela améliore l'interprétabilité.

3.5 Critères d'information

La vraisemblance peut également servir d'élément de base pour la comparaison des modèles : plus $\ell(\widehat{\boldsymbol{\theta}})$ est grand, meilleure est l'adéquation. Cependant, la vraisemblance ne tient pas compte de la complexité du modèle dans le sens où des modèles plus complexes avec plus de paramètres conduisent à une vraisemblance plus élevée. Cela ne pose pas de problème pour la comparaison de modèles emboîtés à l'aide du test du rapport de vraisemblance, car nous ne tenons compte que de l'amélioration relative de l'adéquation. Il existe un risque de **surajustement** si l'on ne tient compte que de la vraisemblance d'un modèle.

Les critères d'information combinent la log vraisemblance, qui mesure l'adéquation du modèle aux données, avec une pénalité pour le nombre de paramètres. Les plus fréquents

sont les critères d'information d'Akaike (AIC) et bayésien (BIC),

$$\begin{aligned}\mathsf{AIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + 2p \\ \mathsf{BIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + p\ln(n),\end{aligned}$$

où p dénote le nombre de paramètres du modèle. Le plus petit la valeur du critère d'information, le meilleur le modèle.

Notez que les critères d'information ne constituent pas des tests d'hypothèse formels sur les paramètres, mais qu'ils peuvent être utilisés pour comparer des modèles non imbriqués (mais ils sont alors très imprécis!) Ces outils fonctionnent sous des conditions de régularité et les critères d'information estimés sont assez bruyants, de sorte que les comparaisons pour les modèles non emboîtés sont hasardeuses bien que populaires. Si nous voulons comparer la vraisemblance de différents modèles de probabilité, nous devons nous assurer qu'ils incluent une constante de normalisation[5]. Le BIC est plus strict que le AIC, car sa pénalité augmente avec la taille de l'échantillon, ce qui permet de sélectionner des modèles plus parsimonieux. Le BIC est un critère **convergent**, ce qui signifie qu'il choisira le vrai modèle parmi un ensemble de modèles avec une probabilité de 1 lorsque $n \to \infty$ si ce dernier fait partie du catalogue de modèles à comparer. En pratique, cela présente peu d'intérêt si l'on suppose que tous les modèles sont des approximations de la réalité (il est peu probable que le vrai modèle soit inclus dans ceux que nous considérons). Pour sa part, AIC sélectionne souvent des modèles trop compliqués dans les grands échantillons, alors que BIC choisit des modèles trop simples.

Une mise en garde s'impose: s'il est possible de comparer des modèles de régression non emboîtés à l'aide de critères d'information, ceux-ci ne peuvent être utilisés que lorsque la variable de réponse est la même. Vous pouvez comparer une régression de Poisson avec une régression linéaire pour une réponse Y en utilisant des critères d'information à condition d'inclure toutes les constantes de normalisation dans votre modèle. Les logiciels omettent souvent les termes constants; cela n'a pas d'impact lorsque vous comparez des modèles avec les mêmes facteurs constants, mais cela a de l'importance lorsque ceux-ci diffèrent. Cependant, **on ne peut pas** les comparer à un modèle log-linéaire avec une réponse $\ln(Y)$. Les comparaisons entre les modèles log-linéaires et linéaires ne sont valables que si vous utilisez la vraisemblance de Box–Cox, car elle inclut le jacobien de la transformation.

[5]Les logiciels enlèvent parfois les termes ou constantes qui ne sont pas des fonctions des paramètres.

4 Régression linéaire

4.1 Introduction

Le modèle de régression linéaire, ou modèle linéaire, est l'un des outils les plus polyvalents pour l'inférence statistique. La régression linéaire est principalement utilisée pour évaluer les effets des variables explicatives (souvent l'effet d'une manipulation ou d'un traitement dans un cadre expérimental) sur la moyenne d'une variable réponse continue, ou pour la prédiction. Un modèle linéaire est un modèle qui décrit la moyenne d'une **variable réponse** continue Y_i d'un échantillon aléatoire de taille n comme **fonction linéaire** des **variables explicatives** (également appelés prédicteurs, régresseurs ou covariables) $X_1, \ldots, X_p$.

Dénotons par Y_i la valeur de Y pour le sujet i, et X_{ij} la valeur de la je variable explicative du sujet i.

$$\underbrace{\mathsf{E}(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i)}_{\text{moyenne conditionnelle}} = \mu_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{\substack{\text{combinaison linéaire (somme pondérée)} \\ \text{de variables explicatives}}} \equiv \mathbf{x}_i \boldsymbol{\beta}. \tag{4.1}$$

où $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ est un vecteur ligne de taille $(p+1)$ contenant les variables explicatives de l'observation i et $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ est un vecteur colonne de longueur $p+1$ contenant les coefficients de la moyenne. Le fait que la moyenne est conditionnelle aux valeurs de $\mathbf{X}$ implique simplement que l'on considère les régresseurs comme constant, ou connus à l'avance. Les coefficients $\boldsymbol{\beta}$ sont les mêmes pour toutes les observations, mais le vecteurs de variables explicatives $\mathbf{x}_i$ peut différer d'une observation à l'autre. Le modèle est **linéaire** en $\beta_0, \ldots, \beta_p$, pas nécessairement dans les variables explicatives.

Pour simplifier la notation, nous regroupons les observations dans un vecteur n $\boldsymbol{Y}$ et les explications dans une matrice $n \times (p+1)$ $\mathbf{X}$ en concaténant une colonne de uns et les vecteurs de colonnes p $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, chacun contenant les n observations des explications respectives. La matrice $\mathbf{X}$ est appelée **matrice du modèle** (ou parfois matrice de devis dans un contexte expérimental), et sa $i$ème ligne est $\mathbf{x}_i$.

En supposant que la variable réponse provient d'une famille de localisation, nous pouvons réécrire le modèle linéaire en termes de la moyenne plus un aléa,

$$\underset{\text{observation}}{Y_i} = \underset{\text{moyenne } \mu_i}{\mathbf{x}_i\boldsymbol{\beta}} + \underset{\text{aléa}}{\varepsilon_i},$$

où ε_i est le terme spécifique à l'observation i. On assume que les aléas $\varepsilon_1, \ldots \varepsilon_n$ sont indépendants et identiquement distribués, avec $\mathsf{E}(\varepsilon_i \mid \mathbf{x}_i) = 0$ et $\mathsf{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2$. On fixe l'espérance de l'aléa à zéro car on postule qu'il n'y a pas d'erreur systématique. La variance σ^2 sert à tenir compte du fait qu'aucune relation linéaire exacte ne lie $\mathbf{x}_i$ et Y_i, ou que les mesures de Y_i sont variables.

Le modèle linéaire normal ou gaussien spécifie que les réponses suivent une loi normale, avec $Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \sim \mathsf{normale}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$. La loi normale est une famille de localisation, de sorte que $Y \sim \mathsf{normale}(\mu, \sigma^2)$ équivaut à la décomposition additive $\mu + \varepsilon$ pour $\varepsilon \sim \mathsf{normale}(0, \sigma^2)$.

4.1.1 Exemples

Considérons quelques exemples de jeux de données qui serviront à illustrer les méthodes par la suite.

Exemple 4.1 (Cohérence de descriptions de produits)**.** L'étude 1 de Lee et Choi (2019) (base de données `LC19_S1`, paquet `hecedsm`) considère l'impact sur la perception d'un produit de la divergence entre la description textuelle et l'image. Dans leur première expérience, un paquet de six brosses à dents est vendu, mais l'image montre soit un paquet de six, soit une seule). Les auteurs ont également mesuré la familiarité préalable avec la marque de l'article. Les $n = 96$ participants ont été recrutés à l'aide d'un panel en ligne. Nous pourrions ajuster un modèle linéaire pour le score moyen d'évaluation du produit, `prodeval`, en fonction de la familiarité de la marque `familiarity`, un nombre entier allant de 1 à 7, et une variable binaire pour le facteur expérimental `consistency`, codé `0` pour des descriptions d'image/texte cohérentes et `1` si elles sont incohérentes. La matrice du modèle qui en résulte est alors de dimension 96×3. La réponse `prodeval` est fortement discrétisée.

```
data(LC19_S1, package = "hecedsm")
modmat <- model.matrix( # Matrice du modèle
     ~ familiarity + consistency,
     data = LC19_S1)
tail(modmat, n = 5L) # Imprimer les premières 5 lignes
```

```
#>    (Intercept) familiarity consistencyinconsistent
#> 92           1           6                       1
#> 93           1           4                       1
#> 94           1           7                       1
#> 95           1           7                       1
#> 96           1           7                       1
dim(modmat) # dimension de la matrice du modèle
#> [1] 96  3
```

Exemple 4.2 (Méthodes d'apprentissage de compréhension de lecture)**.** La base de données `BSJ92` du paquet `hecedsm` contient les résultats d'une expérience de Baumann, Seifert-Kessell, et Jones (1992) sur l'efficacité de différentes stratégies de lecture sur la compréhension d'enfants.

> Soixante-six élèves de quatrième année ont été assignés au hasard à l'un des trois groupes expérimentaux suivants : (a) un groupe « Think-Aloud » (TA), dans lequel les élèves ont appris diverses stratégies de contrôle de la compréhension pour la lecture d'histoires (par exemple : auto-questionnement, prédiction, relecture) par le biais de la réflexion à haute voix; (b) un groupe lecture dirigée-activité de réflexion (DRTA), dans lequel les élèves ont appris une stratégie de prédiction-vérification pour lire et répondre aux histoires; ou (c) un groupe activité de lecture dirigée (DRA), un groupe contrôle dans lequel les élèves se sont engagés dans une lecture guidée non interactive d'histoires.

Les variables d'intérêt sont `group`, le facteur pour le groupe expérimental, soit `DRTA`, `TA` et `DR` ainsi que les variables numériques `pretest1` et `posttest1`, qui donnent le score (sur 16) sur le test pré-expérience pour la tâche de détection des erreurs.

Les données sont balancées puisqu'il y a 22 observations dans chacun des trois sous-groupes. Les chercheurs ont appliqué une série de trois évaluations: le test 1 de détection d'erreurs, le test 2 consistant en un questionnaire de suivi de compréhension, et le test 3 standardisé *Degrees of Reading Power*). Les tests 1 et 2 ont été administrés à la fois avant et après l'intervention: cela nous permet d'établir l'amélioration moyenne de l'élève en ajoutant le résultat du test pré-intervention comme covariable. Les tests 1 étaient sur 16, mais celui administré après l'expérience a été rendu plus difficile pour éviter les cas d'étudiants obtenant des scores presque complets. La corrélation entre le pré-test et le post-test 1 est $(\widehat{\rho}_1 = 0.57)$, beaucoup plus forte que celle du second test $(\widehat{\rho}_2 = 0.21)$.

Exemple 4.3 (Discrimination salariale dans un collège américain)**.** On s'intéresse à la discrimination salariale dans un collège américain, au sein duquel une étude a été réalisée

pour investiguer s'il existait des inégalités salariales entre hommes et femmes. Le jeu de données `college` contient les variables suivantes:

- `salaire`: salaire de professeurs pendant l'année académique 2008–2009 (en milliers de dollars USD).
- `echelon`: échelon académique, soit adjoint (`adjoint`), aggrégé (`aggrege`) ou titulaire (`titulaire`).
- `domaine`: variable catégorielle indiquant le champ d'expertise du professeur, soit appliqué (`applique`) ou théorique (`theorique`).
- `sexe`: indicateur binaire pour le sexe, `homme` ou `femme`.
- `service`: nombre d'années de service.
- `annees`: nombre d'années depuis l'obtention du doctorat.

Exemple 4.4 (Suggestion de montants de dons)**.** L'étude 1 de Moon et VanEpps (2023) (données `MV23_S1`, paquet `hecedsm`) porte sur la proportion de donateurs à un organisme de charité. Les participants au panel en ligne avaient la possibilité de gagner 25$ et de faire don d'une partie de cette somme à l'organisme de leur choix. Les données fournies incluent uniquement les personnes qui n'ont pas dépassé ce montant et qui ont indiqué avoir fait un don d'un montant non nul.

4.1.2 Analyse exploratoire des données

L'analyse exploratoire des données est une procédure itérative par laquelle nous interrogeons les données, en utilisant des informations auxiliaires, des statistiques descriptives et des graphiques, afin de mieux informer notre modélisation.

Elle est utile pour mieux comprendre les caractéristiques des données (plan d'échantillonnage, valeurs manquantes, valeurs aberrantes), la nature des observations, qu'il s'agisse de variables réponse ou explicatives et les interrelations entre variables.

Voir le Chapitre 11 de Alexander (2023) pour des exemples. En particulier, il convient de vérifier

- que les variables catégorielles sont adéquatement traitées comme des facteurs (`factor`).
- que les valeurs manquantes sont adéquatement déclarées comme telles (code d'erreur, 999, etc.)
- s'il ne vaudrait mieux pas retirer certaines variables explicatives avec beaucoup de valeurs manquantes.
- s'il ne vaudrait mieux pas fusionner des modalités de variables catégorielles si le nombre d'observation par modalité est trop faible.

- qu'il n'y a pas de variable explicative dérivée de la variable réponse
- que le sous-ensemble des observations employé pour l'analyse statistique est adéquat.
- qu'il n'y a pas d'anomalies ou de valeurs aberrantes (par ex., 999 pour valeurs manquantes) qui viendraient fausser les résultats.

Exemple 4.5 (Analyse exploratoire des données `college`)**.** Une analyse exploratoire des données est de mise avant d'ébaucher un modèle. Si le salaire augmente au fil des ans, on voit que l'hétérogénéité change en fonction de l'échelon et qu'il y a une relation claire entre ce dernier et le nombre d'années de service (les professeurs n'étant éligibles à des promotions qu'après un certain nombre d'années). Les professeurs adjoints qui ne sont pas promus sont généralement mis à la porte, aussi il y a moins d'occasions pour que les salaires varient sur cette échelle.

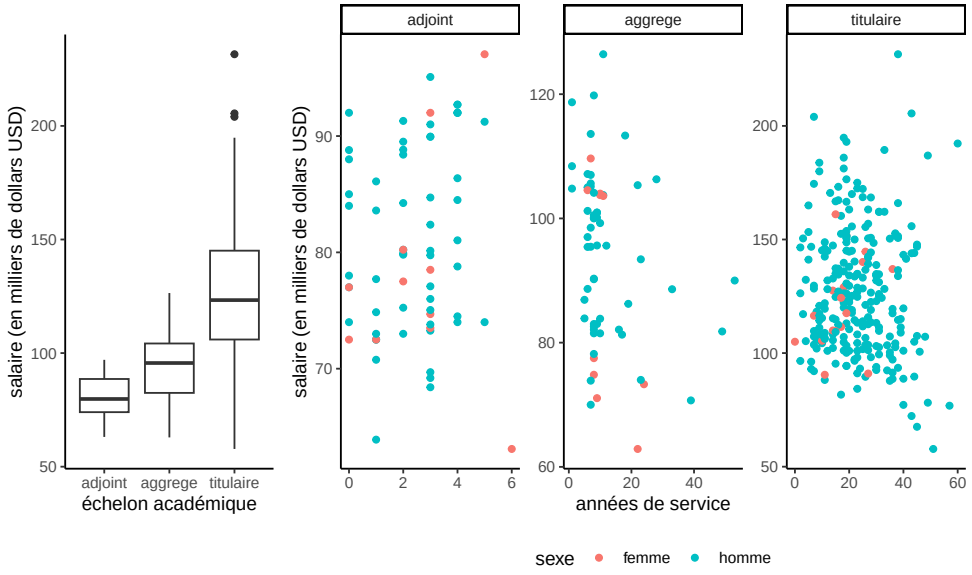

Figure 4.1: Analyse exploratoire des données `college`: répartition des salaires en fonction de l'échelon et du nombre d'années de service

Ainsi, le salaire augmente avec les années, mais la variabilité croît également. Les professeurs adjoints qui ne sont pas promus sont généralement mis à la porte, aussi il y a moins d'occasions pour que les salaires varient sur cette échelle. Il y a peu de femmes dans l'échantillon: moins d'information signifie moins de puissance pour détecter de petites différences de salaire. Si on fait un tableau de contingence de l'échelon et du sexe, on peut calculer la proportion relative homme/femme dans chaque échelon: 16% des profs

adjoints, 16% pour les aggrégés, mais seulement 7% des titulaires alors que ces derniers sont mieux payés en moyenne.

Tableau 4.1: Tableau de contingence donnant le nombre de professeurs du collège par sexe et par échelon académique.

	adjoint	aggrege	titulaire
femme	11	10	18
homme	56	54	248

Plusieurs des variables explicatives potentielles des données `college` sont cat/gorielles (`echelon`, `sexe`, `discipline`), les deux dernières étant binaires. Les variables numériques `annees` et `service` sont fortement corrélées, avec une corrélation linéaire de 0.91.

Exemple 4.6 (Analyse exploratoire et données manquantes)**.** Il convient de vérifier pour les données de Moon et VanEpps (2023) que la description de la collecte coïncide avec la structure. Puisque les personnes qui n'ont pas donné ne remplissent pas le champ pour le montant, ce dernier indique une valeur manquante. Tous les montants des dons sont entre 0.25$ et 25$.

```
data(MV23_S1, package = "hecedsm")
str(MV23_S1)
#> tibble [869 x 4] (S3: tbl_df/tbl/data.frame)
#>  $ before   : int [1:869] 0 1 0 1 1 1 1 0 1 0 ...
#>  $ donate   : int [1:869] 0 0 0 1 1 0 1 0 0 1 ...
#>  $ condition: Factor w/ 2 levels "open-ended","quantity": 1 1 1 1 2 2 2 1 1 1 ...
#>  $ amount   : num [1:869] NA NA NA 10 5 NA 20 NA NA 25 ...
summary(MV23_S1)
#>      before          donate          condition       amount
#>  Min.   :0.000   Min.   :0.00   open-ended:407   Min.   : 0.2
#>  1st Qu.:0.000   1st Qu.:0.00   quantity  :462   1st Qu.: 5.0
#>  Median :1.000   Median :1.00                    Median :10.0
#>  Mean   :0.596   Mean   :0.73                    Mean   :10.7
#>  3rd Qu.:1.000   3rd Qu.:1.00                    3rd Qu.:15.0
#>  Max.   :1.000   Max.   :1.00                    Max.   :25.0
#>  NA's   :1                                       NA's   :235
```

Si nous incluons `amount` comme variable réponse dans un modèle de régression, les 235 observations manquantes seront supprimées par défaut. Cela ne pose pas de problème si

nous voulons comparer le montant moyen des personnes qui ont fait un don, mais dans le cas contraire, nous devons transformer les `NA` en zéros. La variable `donate` ne doit pas être incluse comme variable explicative dans le modèle, car elle permet de prédire exactement les personnes qui n'ont pas donné.

4.1.3 Spécification du modèle pour la moyenne

La première étape d'une analyse consiste à décider quelles variables explicatives doivent être ajoutées à l'équation de la moyenne, et sous quelle forme. Les modèles ne sont que des approximations de la réalité; la section 2.1 de Venables (2000) affirme que, si nous pensons que la véritable fonction moyenne reliant les variables explicatives $\boldsymbol{X}$ et la réponse Y est de la forme $\mathsf{E}(Y \mid \boldsymbol{X}) = f(\boldsymbol{X})$ pour f suffisamment lisse, alors le modèle linéaire est une approximation du premier ordre. À des fins d'interprétation, il est logique de centrer sur la moyenne toute variable explicative continue, car cela facilite l'interprétation.

Dans un cadre expérimental, où la condition expérimentale est attribué de manière aléatoire, nous pouvons directement comparer les différents traitements et tirer des conclusions causales (puisque toutes les autres choses sont égales en moyenne constantes, toute différence détectable est due en moyenne à notre manipulation). Bien que nous nous abstenions généralement d'inclure d'autres variables explicatives afin de préserver la simplicité du modèle, il peut néanmoins être utile de prendre en compte certaines variables concomitantes qui expliquent une partie de la variabilité afin de filtrer le bruit de fond et d'augmenter la puissance de l'étude. Par exemple, pour les données de Baumann, Seifert-Kessell, et Jones (1992), l'objectif est de comparer les scores moyens en fonction de la méthode d'enseignement, nous inclurions `group`. Dans cet exemple, il serait également logique d'inclure le résultat `pretest1` en tant qu'élément explicatif pour `posttest1`. De cette façon, nous modéliserons la différence moyenne d'amélioration entre le pré-test et le post-test plutôt que le résultat final.

Dans un contexte observationnel, les participants dans différents groupes ont des caractéristiques différentes et nous devons donc tenir compte de ces différences. Les modèles linéaires utilisés en économie et en finance contiennent souvent des variables de contrôle au modèle pour tenir compte des différences potentielles dues aux variables sociodémographiques (âge, revenu, etc.) qui seraient corrélées à l'appartenance aux groupes. Tout test de coefficients ne prendrait en compte que la corrélation entre le résultat Y et le facteur explicatif postulé d'intérêt.

4.2 Interprétation des coefficients

La spécification de la moyenne est

$$\mathsf{E}(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

L'**ordonnée à l'origine** β_0 est la **valeur moyenne de** Y lorsque toutes les variables explicatives du modèles sont nulles, soit $\boldsymbol{x}_i = \mathbf{0}_p$.

$$\begin{aligned}\beta_0 &= \mathsf{E}(Y \mid X_1 = 0, X_2 = 0, \ldots, X_p = 0) \\ &= \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \cdots + \beta_p \times 0\end{aligned}$$

Bien sur, il se peut que cette interprétation n'ait aucun sens dans le contexte étudié. Centrer les variables explicatives numériques (pour que leurs moyennes soit zéro) permet de rendre l'ordonnée à l'origine plus interprétable.

En régression linéaire, le paramètre β_j mesure l'effet de la variable X_j sur la variable Y une fois que l'on tient compte des effets des autres variables explicatives. Pour chaque augmentation d'une unité de X_j, la réponse Y augmente en moyenne de β_j lorsque les autres variables demeurent inchangées,

$$\begin{aligned}\beta_j &= \mathsf{E}(Y \mid X_j = x_j + 1, \boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) - \mathsf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{\substack{k=1 \\ k \neq j}}^{p} \beta_k x_k + \beta_j (x_j + 1) - \sum_{k=1}^{p} \beta_k x_k\end{aligned}$$

Définition 4.1 (Effet marginal)**.** On définit l'effet marginal comme la dérivée première de la moyenne conditionnelle par rapport à X_j, soit

$$\text{effet marginal de } X_j = \frac{\partial \mathsf{E}(Y \mid \boldsymbol{X})}{\partial X_j}.$$

Le coefficient β_j est aussi l'*effet marginal* de la variable X_j.

Les variables indicatrices, qui prennent typiquement des valeurs de -1, 0 et 1, servent à indiquer l'appartenance aux différentes modalités d'une variable catégorielle. Par exemple, pour une variable indicatrice binaire, nous pouvons créer une colonne dont les entrées sont 1 pour le groupe de traitement et 0 pour le groupe de contrôle.

Exemple 4.7 (Modèle linéaire avec une seule variable binaire)**.** Considérons par exemple un modèle linéaire pour les données de Moon et VanEpps (2023) qui inclut le montant (`amount`) (en dollars, de 0 pour les personnes qui n'ont pas fait de don, jusqu'à 25 dollars).

L'équation du modèle linéaire simple qui inclut la variable binaire `condition` est

$$\begin{aligned}\mathsf{E}(\texttt{amount} \mid \texttt{condition}) &= \beta_0 + \beta_1 \mathbf{1}_{\texttt{condition=quantity}}. \\ &= \begin{cases} \beta_0, & \texttt{condition} = 0, \\ \beta_0 + \beta_1 & \texttt{condition} = 1. \end{cases}\end{aligned}$$

Soit μ_0 l'espérance du montant pour le groupe contrôle (`open-ended`) et μ_1 celui des participants du groupe de traitement (`quantity`). Un modèle linéaire qui ne contient qu'une variable binaire X comme régresseur revient à spécifier une moyenne différente pour chacun des deux groupes. L'ordonnée à l'origine β_0 est la moyenne du groupe contrôle. La moyenne du groupe traitement (`quantity`) est $\beta_0 + \beta_1 = \mu_1$ et donc $\beta_1 = \mu_1 - \mu_0$ est la différence du montant moyen de dons entre le groupe `open-ended` et le groupe `quantity`. Cette paramétrisation est commode si on veut tester s'il y a une différence moyenne entre les deux groupes, puisque cette hypothèse nulle correspond à $\mathscr{H}_0 : \beta_1 = 0$.

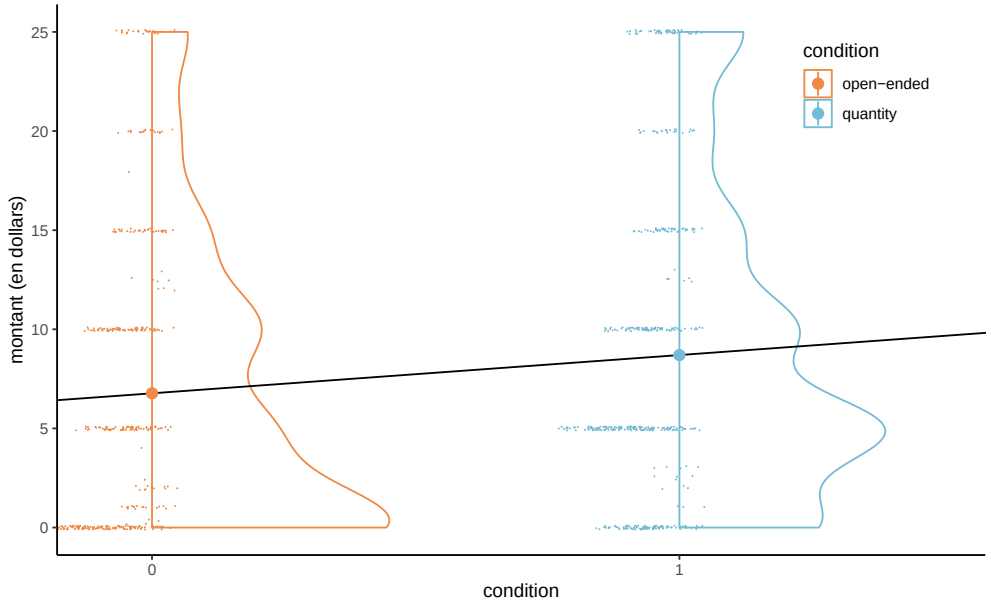

Figure 4.2: Modèle linéaire simple pour les données `MV23_S1` avec `condition` comme variable explicative binaire, avec nuage de points décalés et un diagramme en demi-violin. Les cercles indiquent les moyennes de l'échantillon.

Même si le modèle linéaire définit une droite, cette dernière ne peut être évaluée qu'à 0 ou 1; la Figure 4.2 montre cette droite avec en plus un nuage de points des montants, décalés horizontalement, et de la densité pour chaque condition. Le point coloré indique la moyenne empirique, qui correspond aux estimations.

Même s'il est clair que les données sont fortement discrétisées avec beaucoup de doublons et de zéros, l'échantillon a une taille de 869 observations, donc les conclusions quant aux moyennes de groupe seront fiables.

Considérons des variables catégorielles avec $K > 2$ niveaux, qui dans **R** sont de la classe `factor`. La paramétrisation par défaut des facteurs se fait en termes de contraste de traitement: le niveau de référence du facteur (par défaut, la première valeur dans l'ordre alphanumérique) sera traité comme la catégorie de référence et assimilé à l'ordonnée à l'origine. Le logiciel créera alors un ensemble de $K-1$ variables indicatrices pour un facteur à K niveaux, chacune d'entre elles ayant un pour la catégorie représentée et zéro dans le cas contraire.

Exemple 4.8 (Codage binaire pour les variables catégorielles)**.** Considérons l'étude de Baumann, Seifert-Kessell, et Jones (1992) et la seule variable `group`. Les données sont classées par groupe : les 22 premières observations concernent le groupe `DR`, les 22 suivantes le groupe `DRTA` et les 22 dernières le groupe `TA`. Si nous ajustons un modèle avec `groupe` comme variable catégorielle

```
data(BSJ92, package = "hecedsm")
class(BSJ92$group) # Vérifier que group est un facteur
#> [1] "factor"
levels(BSJ92$group) # première valeur est la catégorie de référence
#> [1] "DR"   "DRTA" "TA"
# Imprimer trois lignes de la matrice du modèle
# (trois enfants de groupes différents)
model.matrix(~ group, data = BSJ92)[c(1,23,47),]
#>    (Intercept) groupDRTA groupTA
#> 1            1         0       0
#> 23           1         1       0
#> 47           1         0       1
# Comparer avec les niveaux des facteurs
BSJ92$group[c(1,23,47)]
#> [1] DR   DRTA TA
#> Levels: DR DRTA TA
```

Si nous ajustons un modèle avec `groupe` comme variable catégorielle, la spécification de la moyenne du modèle est

$$\mathsf{E}(Y \mid \texttt{group}) = \beta_0 + \beta_1 \mathbf{1}_{\texttt{group=DRTA}} + \beta_2 \mathbf{1}_{\texttt{group=TA}}.$$

Puisque la variable `group` est catégorielle avec $K = 3$ niveaux, il nous faut mettre $K - 1 = 2$ variables indicatrices.

Avec la paramétrisation en termes de **traitements** (option par défaut), on obtient

- $\mathbf{1}_{\texttt{group=DRTA}} = 1$ si `group=DRTA` et zéro sinon,
- $\mathbf{1}_{\texttt{group=TA}} = 1$ si `group=TA` et zéro sinon.

Étant donné que le modèle comprend une ordonnée à l'origine et que le modèle décrit en fin de compte trois moyennes de groupe, nous n'avons besoin que de deux variables supplémentaires. Avec la paramétrisation en termes de **traitements**, la moyenne du groupe de référence est l'ordonnée à l'origine. Si `group=DR` (référence), les deux variables indicatrices binaires `groupDRTA` et `groupTA` sont nulles. La moyenne de chaque groupe est

- $\mu_{\texttt{DR}} = \beta_0$,
- $\mu_{\texttt{DRTA}} = \beta_0 + \beta_1$ et
- $\mu_{\texttt{TA}} = \beta_0 + \beta_2$.

Ainsi, β_1 est la différence de moyenne entre les groupes `DRTA` et`DR`, et de la même façon $\beta_2 = \mu_{\texttt{TA}} - \mu_{\texttt{DR}}$.

Remarque 4.1 (Contrainte de somme nulle). La paramétrisation discutée ci-dessus, qui est la valeur par défaut de la fonction `lm`, n'est pas la seule disponible. Plutôt que de comparer la moyenne de chaque groupe avec celle d'une catégorie de référence, la paramétrisation par défaut pour les modèles d'analyse de la variance est en termes de contraintes de somme nulle pour les coefficients, où l'ordonnée à l'origine est la moyenne équi-pondérée de chaque groupe, et les paramètres $\beta_1, \ldots, \beta_{K-1}$ sont des différences par rapport à cette moyenne.

```
model.matrix(
    ~ group,
    data = BSJ92,
    contrasts.arg = list(group = "contr.sum"))
```

Tableau 4.2: Paramétrisation des variables indicatrices pour la contrainte de somme nulle pour une variable catégorielle.

	(Intercept)	group1	group2
DR	1	1	0
DRTA	1	0	1

Tableau 4.2: Paramétrisation des variables indicatrices pour la contrainte de somme nulle pour une variable catégorielle.

	(Intercept)	group1	group2
TA	1	-1	-1

Dans la contrainte de somme nulle, nous obtenons à nouveau deux variables indicatrices, `group1` et `group2`, ainsi que l'ordonnée à l'origine. La valeur de `group1` est 1 si `group=DR`, 0 si `group=DRTA` et -1 si `group=TA`. ous trouvons $\mu_{\texttt{DR}} = \beta_0 + \beta_1$, $\mu_{\texttt{DRTA}} = \beta_0 + \beta_2$ et $\mu_{\texttt{TA}} = \beta_0 - \beta_1 - \beta_2$. Quelques manipulations algébriques révèlent que $\beta_0 = (\mu_{\texttt{DR}} + \mu_{\texttt{DRTA}} + \mu_{\texttt{TA}})/3$, l'espérance équipondérée des différents niveaux. De manière générale, l'ordonnée à l'origine moins la somme de tous les autres coefficients liés aux facteurs.

En supprimant l'ordonnée à l'origine, on pourrait inclure trois variables indicatrices pour chaque niveau d'un facteur et chaque paramètre correspondrait alors à la moyenne. Ce n'est pas recommandé dans **R** car le logiciel traite différemment les modèles sans ordonnée à l'origine et certains résultats seront absurdes (par exemple, le coefficient de détermination sera erroné).

Exemple 4.9 (Interprétation des coefficients)**.** On considère un modèle de régression pour les données `college` qui inclut le sexe, l'échelon académique, le nombre d'années de service et le domaine d'expertise (appliquée ou théorique).

Le modèle linéaire postulé s'écrit

$$\begin{aligned}\texttt{salaire} = \beta_0 &+ \beta_1 \mathbf{1}_{\texttt{sexe=femme}} + \beta_2 \mathbf{1}_{\texttt{domaine=theorique}} \\ &+ \beta_3 \mathbf{1}_{\texttt{echelon=aggrege}} + \beta_4 \mathbf{1}_{\texttt{echelon=titulaire}} + \beta_5 \texttt{service} + \varepsilon.\end{aligned}$$

Tableau 4.3: Estimations des coefficients du modèle linéaire pour les données `college` (en dollars USD, arrondis à l'unité).

$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
86596	-4771	-13473	14560	49160	-89

L'interprétation des coefficients est la suivante:

- L'ordonnée à l'origine β_0 correspond au salaire moyen d'un professeur adjoint (un homme) qui vient de compléter ses études et qui travaille dans un domaine appliqué: on estime ce salaire à $\widehat{\beta}_0 = 86596$ dollars.
- toutes choses étant égales par ailleurs (même domaine, échelon et années depuis le dernier diplôme), l'écart de salaire entre un homme et un femme est estimé à $\widehat{\beta}_1 = -4771$ dollars.
- *ceteris paribus*, un(e) professeur(e) qui oeuvre dans un domaine théorique gagne β_2 dollars de plus qu'une personne du même sexe dans un domaine appliqué; on estime cette différence à -13473 dollars.
- *ceteris paribus*, la différence moyenne de salaire entre professeurs adjoints et aggrégés est estimée à $\widehat{\beta}_3 = 14560$ dollars.
- *ceteris paribus*, la différence moyenne de salaire entre professeurs adjoints et titulaires est de $\widehat{\beta}_4 = 49160$ dollars.
- au sein d'un même échelon, chaque année supplémentaire de service mène à une augmentation de salaire annuelle moyenne de $\widehat{\beta}_5 = -89$ dollars.

Remarque 4.2 (Polynômes). Il n'est pas toujours possible de fixer la valeur des autres colonnes de $\mathbf{X}$ si plusieurs colonnes contiennent des transformations ou des fonctions d'une même variable explicative. Par exemple, on pourrait par exemple considérer un polynôme d'ordre k (ordinairement, on prendre $k \leq 3$),

$$\mathsf{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k.$$

Si l'on inclut un terme d'ordre k, x^k, il faut **toujours** inclure les termes d'ordre inférieur $1, x, \ldots, x^{k-1}$ pour l'interprétabilité du modèle résultant (autrement, cela revient à choisir un polynôme en imposant que certains coefficients soient zéros). L'interprétation des effets des covariables nonlinéaires (même polynomiaux) est complexe parce qu'on ne peut pas « fixer la valeur des autres variables »: l'effet d'une augmentation d'une unité de x *dépend de la valeur de cette dernière*. L'effet marginal de x est $\beta_1 + \sum_{j=1}^{k-1} j\beta_{j+1}x^j$.

L'utilisation de polynôme, plus flexibles, n'est généralement pas recommendée car ces derniers se généralisent mal hors de l'étendue observée des données. L'utilisation de splines avec une pénalité sur les coefficients, avec des modèles additifs, offre plus de flexibilité.

Exemple 4.10 (Modèle quadratique pour les données automobile)**.** Considérons un modèle de régression linéaire pour l'autonomie d'essence en fonction de la puissance du moteur pour différentes voitures dont les caractéristiques sont données dans le jeu de données `automobiles`. Le modèle postulé incluant un terme quadratique est

$$\texttt{autonomie}_i = \beta_0 + \beta_1 \texttt{puissance}_i + \beta_2 \texttt{puissance}_i^2 + \varepsilon_i$$

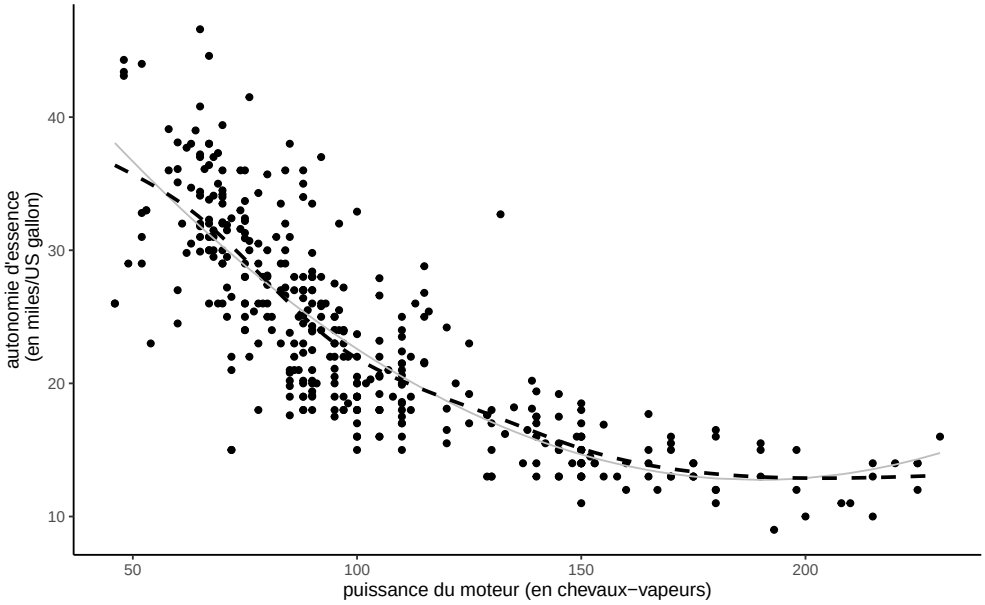

Figure 4.3: Modèle de régression avec terme quadratique pour la puissance (gris), versus spline cubique pénalisée (ligne traitillée).

Afin de comparer l'ajustement du modèle quadratique, on peut inclure également la droite ajustée du modèle de régression simple qui n'inclut que puissance.

À vue d'oeil, l'ajustement quadratique est bon: nous verrons plus tard à l'aide de test si une simple droite aurait été suffisante. On voit aussi dans la Figure 4.3 que l'autonomie d'essence décroît rapidement quand la puissance croît entre 0 et 189.35, mais semble remonter légèrement par la suite pour les voitures qui un moteur de plus de 200 chevaux-vapeurs, ce que le modèle quadratique capture. Prenez garde en revanche à l'extrapolation là où vous n'avez pas de données (comme l'illustre remarquablement bien le modèle cubique de Hassett pour le nombre de cas quotidiens de coronavirus).

La représentation graphique du modèle polynomial de degré 2 présenté dans la Figure 4.3 peut sembler contre-intuitive, mais c'est une projection en 2D d'un plan 3D de coordonnées $\beta_0 + \beta_1 x - y + \beta_2 z = 0$, où $x = \texttt{puissance}$, $z = \texttt{puissance}^2$ et $y = \texttt{autonomie}$. La physique et le bon-sens imposent la contrainte $z = x^2$, et donc les valeurs ajustées vivent sur une courbe dans un sous-espace du plan ajusté, représenté en gris dans la Figure 4.4.

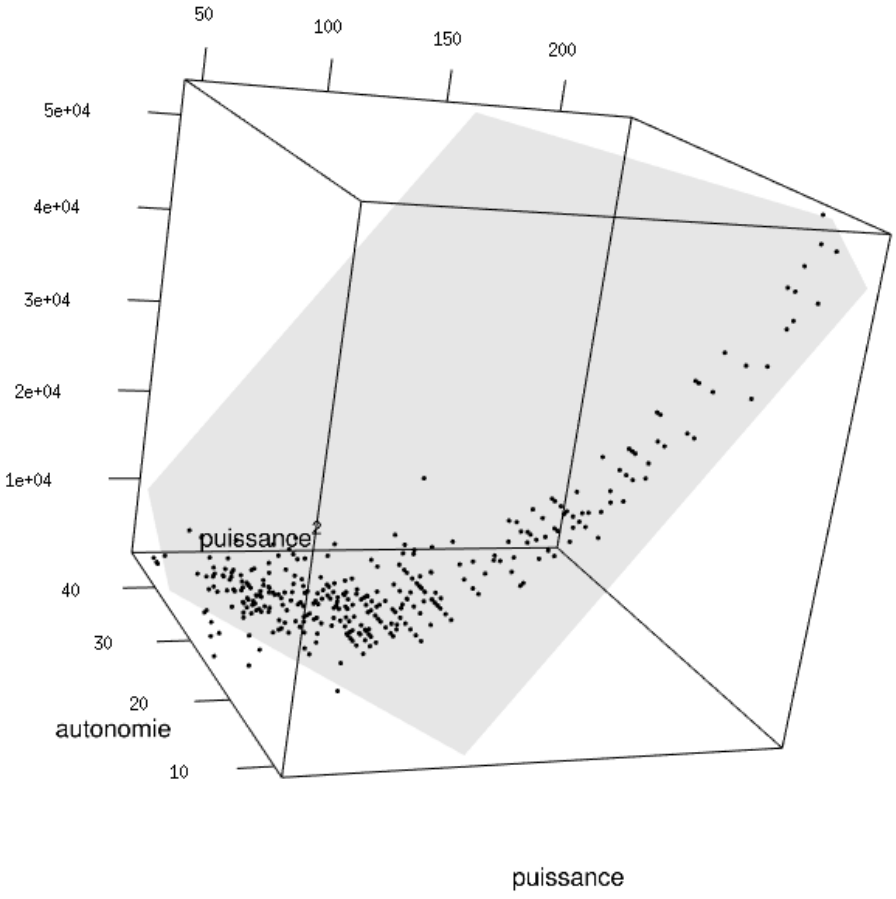

Figure 4.4: Représentation graphique 3D du modèle de régression linéaire pour les données `automobile`.

4.3 Estimation des paramètres

Considérons un échantillon de n observations. On n'observe ni les aléas ε, ni les paramètres β: il est donc impossible de recouvrer les (vrais) coefficients du modèle. Effectivement, le système d'équation spécifié par le modèle linéaire inclut $n + p + 1$ inconnues, mais uniquement n observations. Si on se concentre sur les $p + 1$ paramètres de moyenne et sur la variance σ^2, nous pourrons estimer les paramètres généralement si $n > p+2$, mais cela dépend de la spécification. Une infinité de plans pourraient passer dans le nuage de points; il faut donc choisir la meilleure droite (selon un critère donné). La section aborde

le choix de ce critère et l'estimation des paramètres de la moyenne.

4.3.1 Moindres carrés ordinaires

Soit une matrice de modèle $\mathbf{X}$ et une formulation pour la moyenne avec $\mathsf{E}(Y_i) = \mathbf{x}_i\boldsymbol{\beta}$. Les estimateurs des moindres carrés ordinaires $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ sont les paramètres qui minimisent simultanément la distance euclidienne entre les observations y_i et les **valeurs ajustées** $\widehat{y}_i = \mathbf{x}_i\widehat{\boldsymbol{\beta}}$.

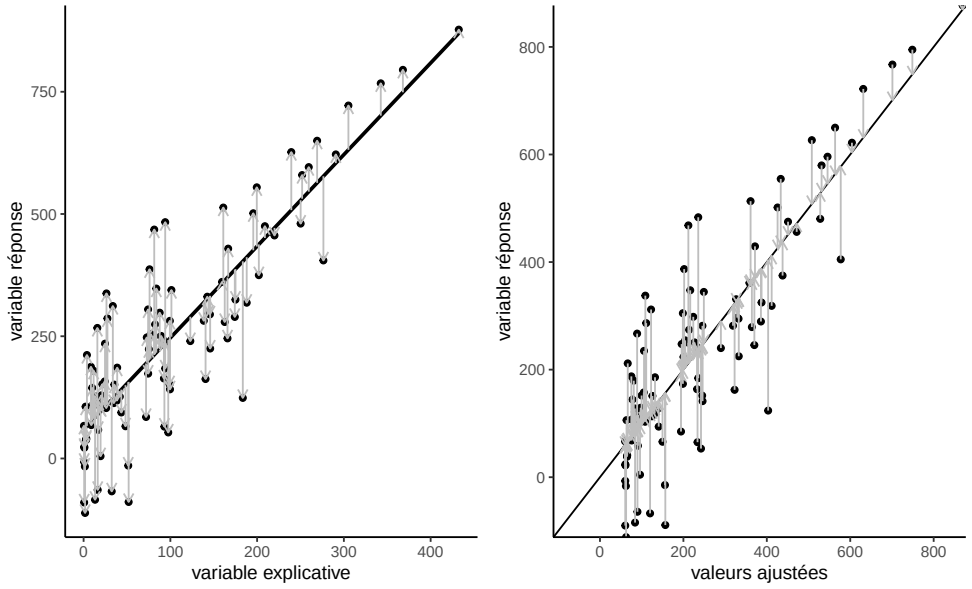

Figure 4.5: Résidus ordinaires e_i (vecteurs verticaux) ajoutés à la droit de régression dans l'espace (x, y) (gauche) et l'ajustement de la variable réponse y_i en fonction des valeurs ajustées $\widehat{y}_i$.

En d'autres mots, les estimateurs des moindres carrés sont la solution du problème d'optimization convexe

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Ce système d'équation a une solution explicite qui est plus facilement exprimée en nota-

tion matricielle. Soit les matrices et vecteurs

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Proposition 4.1 (Moindres carrés ordinaires)**.** *L'estimateur des moindres carrés ordinaires résoud le problème d'optimisation non-contraint*

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

On peut calculer la dérivée première par rapport à $\boldsymbol{\beta}$, égaler à zéro et isoler le maximum pour obtenir une formule explicite pour $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned} \mathbf{0}_n &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})} \\ &= \mathbf{X}^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

*en utilisant la règle de dérivation en chaîne; on peut ainsi distribuer les termes pour obtenir l'*équation normale

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \boldsymbol{y}.$$

Si $\mathbf{X}$ est une matrice de rang p, alors la forme quadratique $\mathbf{X}^\top \mathbf{X}$ est inversible et l'unique solution du problème d'optimisation est

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}.$$

Si le rang de la matrice $\mathbf{X}$ est dimension $n \times (p+1)$ est de rang $p+1$, l'unique solution du problème d'optimisation est

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}. \tag{4.2}$$

*Cet estimateur dit des **moindres carrés ordinaires** (MCO) est explicite; il n'est donc pas nécessaire de procéder à l'optimisation à l'aide d'algorithmes numériques.*

4.3.2 Maximum de vraisemblance

Nous pourrions également envisager l'estimation du maximum de vraisemblance. Proposition 4.2 montre que, en supposant la normalité des aléas, les estimateurs des moindres carrés de β coïncident avec ceux du maximum de vraisemblance.

Proposition 4.2 (Estimation du maximum de vraisemblance du modèle linéaire normal)**.** *Le modèle de régression linéaire spécifie que les observations* $Y_i \sim \mathsf{normale}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ *sont indépendantes. Le modèle linéaire a* $p+2$ *paramètres (*$\boldsymbol{\beta}$ *et* σ^2*) et la log-vraisemblance est, abstraction faite des termes constants,*

$$\ell(\boldsymbol{\beta}, \sigma) \propto -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left\{(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})\right\}^2.$$

Maximiser la log-vraisemblance par rapport à $\boldsymbol{\beta}$ *revient à minimiser la somme du carré des erreurs* $\sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2$*, quelle que soit la valeur de* σ*, et on recouvre* $\widehat{\boldsymbol{\beta}}$*. L'estimateur du maximum de vraisemblance de la variance* $\widehat{\sigma}^2$ *est*

$$\widehat{\sigma}^2 = \mathrm{argmax}_{\sigma^2}\ell(\widehat{\boldsymbol{\beta}}, \sigma^2).$$

La log-vraisemblance profilée de σ^2*, abstraction faite des constantes, est*

$$\ell_{\mathrm{p}}(\sigma^2) \propto -\frac{1}{2}\left\{n\ln\sigma^2 + \frac{1}{\sigma^2}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right\}.$$

En différenciant chaque terme par rapport à σ^2 *et en fixant le gradient à zéro, on obtient*

$$\frac{\partial\ell_{\mathrm{p}}(\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2\sigma^4} = 0$$

On déduit que l'estimateur du maximum de vraisemblance est la moyenne des carrés des résidus,

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})^2 = \frac{\mathsf{SC}_e}{n};\end{aligned}$$

L'estimateur sans biais habituel de σ^2 *calculé par le logiciel est*

$$S^2 = \mathsf{SC}_e/(n-p-1),$$

où le dénominateur est la taille de l'échantillon n *moins le nombre de paramètres de la moyenne* $\boldsymbol{\beta}$*, soit* $p+1$*.*

Remarque 4.3 (Invariance). Une conséquence directe des propriétés des estimateurs du maximum de vraisemblance est que les valeurs ajustées $\widehat{y}_i$ pour deux matrices de modèle $\mathbf{X}_a$ et $\mathbf{X}_b$ sont les mêmes si elles engendrent le même espace linéaire, comme dans Exemple 4.8; seule l' 'interprétation des coefficients change. Si nous incluons une ordonnée à l'origine, nous obtenons le même résultat si les colonnes explicatives sont centrées sur la moyenne.

La valeur de $\widehat{\boldsymbol{\beta}}$ est telle qu'elle maximise la corrélation entre $\boldsymbol{y}$ et $\widehat{\boldsymbol{y}}$. Dans le cas d'une variable catégorielle unique, nous obtiendrons des valeurs ajustées $\widehat{y}$ qui correspondent à la moyenne de l'échantillon de chaque groupe.

Remarque 4.4 (Géométrie). Le vecteur de valeurs ajustées $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}_{\mathbf{X}}\boldsymbol{y}$ est la projection du vecteur réponse $\boldsymbol{y}$ dans l'espace linéaire engendré par les colonnes de $\mathbf{X}$. La matrice chapeau $\mathbf{H}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ est une matrice de projection orthogonale, car $\mathbf{H}_{\mathbf{X}} = \mathbf{H}_{\mathbf{X}}^\top$ et $\mathbf{H}_{\mathbf{X}}\mathbf{H}_{\mathbf{X}} = \mathbf{H}_{\mathbf{X}}$. Ainsi, $\mathbf{H}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$. Puisque le vecteur de résidus ordinaires $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$, qui apparaît dans la somme des erreurs quadratiques, est définie comme $\boldsymbol{y} - \widehat{\boldsymbol{y}}$ et $\widehat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{\beta}$, de simples manipulations algébriques montrent que le produit scalaire entre les résidus ordinaires et les valeurs ajustées est nul, puisque

$$\begin{aligned}
\widehat{\boldsymbol{y}}^\top\boldsymbol{e} &= \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \boldsymbol{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}) \\
&= \boldsymbol{y}^\top\mathbf{H}_{\mathbf{X}}\boldsymbol{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \\
&= 0
\end{aligned}$$

où nous utilisons la définition de $\widehat{\boldsymbol{y}}$ et $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ sur la première ligne, puis on substitut l'estimateur des MCO $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ avant de distribuer les termes du produit. Une dérivation similaire montre que $\mathbf{X}^\top\boldsymbol{e} = \mathbf{0}_{p+1}$. Les résidus ordinaires sont donc orthogonaux à la fois à la matrice du modèle $\mathbf{X}$ et aux valeurs ajustées $\widehat{\boldsymbol{y}}$.

Une conséquence directe de ces résultats est le fait que la corrélation linéaire entre $\boldsymbol{e}$ et $\widehat{\boldsymbol{y}}$ est nulle. Cette propriété servira lors de l'élaboration de diagnostics graphiques.

Puisque le produit scalaire est zéro, la moyenne de $\boldsymbol{e}$ doit être zéro pour autant que $\mathbf{1}_n$ est dans l'espace linéaire engendré par $\mathbf{X}$.

Proposition 4.3 (Matrices d'information pour modèles linéaires normaux.)**.** *Les entrées de*

la matrice d'information observée du modèle linéaire normal sont les suivantes

$$\begin{aligned}
-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \frac{1}{\sigma^2}\frac{\partial \mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2} \\
-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \sigma^2} &= -\frac{\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^4} \\
-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial (\sigma^2)^2} &= -\frac{n}{2\sigma^4} + \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^6}.
\end{aligned}$$

Si on évalue l'information observée aux EMV, on obtient

$$j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}) = \begin{pmatrix} \frac{\mathbf{X}^\top\mathbf{X}}{\widehat{\sigma^2}} & \mathbf{0}_{p+1} \\ \mathbf{0}_{p+1}^\top & \frac{n}{2\widehat{\sigma^4}} \end{pmatrix}$$

puisque $\widehat{\sigma}^2 = \mathsf{SC}_e/n$ *et que les résidus sont orthogonaux à la matrice du modèle. Sachant que* $\mathsf{E}(Y \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$*, la matrice d'information de Fisher est*

$$i(\boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2} & \mathbf{0}_{p+1} \\ \mathbf{0}_{p+1}^\top & \frac{n}{2\sigma^4} \end{pmatrix}$$

Puisque la loi asymptotique de l'estimateur est normale, les EMV de σ^2 *et* $\boldsymbol{\beta}$ *sont asymptotiquement indépendants car leur corrélation asymptotique est nulle.Pourvu que la matrice carrée* $(p+1)$*,* $\mathbf{X}^\top\mathbf{X}$ *soit inversible, la variance asymptotique des estimateurs est* $\mathsf{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ *et* $\mathsf{Var}(\widehat{\sigma}^2) = 2\sigma^4/n$*.*

4.3.3 Ajustement des modèles linéaires à l'aide d'un logiciel

Bien que nous puissions construire la matrice du modèle nous-mêmes et utiliser la formule des moindres carrés de l'Équation 4.2, les routines numériques implémentées dans les logiciels sont préférables car plus stables. La fonction `lm` dans **R** ajuste **les modèles linéaires**, tout comme `glm` avec les arguments par défaut. Les objets de la classe `lm` ont plusieurs méthodes qui vous permettent d'extraire des objets spécifiques des objets `lm`. Par exemple, les fonctions `coef`, `resid`, `fitted`, `model.matrix` renvoient les estimations des coefficients $\widehat{\boldsymbol{\beta}}$, les résidus ordinaires $\boldsymbol{e}$, les valeurs ajustées $\widehat{\boldsymbol{y}}$ et la matrice du modèle $\mathbf{X}$.

```r
data(BSJ92, package = "hecedsm") # charger les données
str(BSJ92) # vérifier que les variables catégorielles sont "factor"
# Ajustement de la régression linéaire
linmod <- lm(posttest1 ~ pretest1 + group,
             data = BSJ92)
est_beta <- coef(linmod) # coefficients (betas)
vcov_beta <- vcov(linmod) # matrice de covariance des betas
summary(linmod) # tableau résumé
beta_ic <- confint(linmod) # IC de Wald pour betas
y_adj <- fitted(linmod) # valeurs ajustées
e <- resid(linmod) # résidus ordinaires

# Vérifier la formule des moindres carrés ordinaires
X <- model.matrix(linmod) # matrice du modèle
y <- college$salary
isTRUE(all.equal(
  c(solve(t(X) %*% X) %*% t(X) %*% y),
  as.numeric(coef(linmod))
))
```

La méthode `summary` est sans doute la plus utile: elle affiche les estimations des paramètres de la moyenne ainsi que leurs erreurs type, les valeurs t pour le test de Wald de l'hypothèse $\mathscr{H}_0 : \beta_i = 0$ et les valeurs-p associées. D'autres statistiques descriptives, portant sur la taille de l'échantillon, les degrés de liberté, etc. sont données au bas du tableau. Notez que la fonction `lm` utilise l'estimateur sans biais de la variance σ^2.

4.4 Coefficient de détermination

Lorsque nous spécifions un modèle, les aléas ε servent à tenir compte du fait qu'aucune relation linéaire exacte ne caractérise les données Une fois que nous avons ajusté un modèle, nous estimons la variance σ^2; on peut alors se demander quelle part de la variance totale de l'échantillon est expliquée par le modèle.

La somme totale des carrés, définie comme la somme des carrés des résidus du modèle à ordonnée à l'origine uniquement, sert de comparaison — le modèle le plus simple que nous puissions trouver impliquerait chaque observation par la moyenne de l'échantillon de la réponse, ce qui donne la variance expliquée $\mathsf{SC}_c = \sum_{i=1}^n (y_i - \overline{y})^2$. Nous pouvons ensuite comparer la variance des données originales avec celle des résidus du modèle avec

la matrice de covariables $\mathbf{X}$, définie comme $\mathsf{SC}_e = \sum_{i=1}^n e_i^2$ avec $e_i = y_i - \widehat{\beta}_0 - \sum_{j=1}^p \widehat{\beta}_j X_j$. Nous définissons le coefficient de détermination R^2, comme suit

$$R^2 = 1 - \frac{\mathsf{SC}_e}{\mathsf{SC}_c} = \frac{\sum_{i=1}^n (y_i - \overline{y})^2 - \sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$

Une autre décomposition montre que $R^2 = \mathsf{cor}^2(\boldsymbol{y}, \widehat{\boldsymbol{y}})$, c'est-à-dire que le coefficient de détermination peut être interprété comme le carré de la corrélation linéaire de Pearson (Définition 1.3) entre la réponse $\boldsymbol{y}$ et les valeurs ajustées $\widehat{\boldsymbol{y}}$.

Il est important de noter que le R^2 n'est pas un critère de qualité de l'ajustement, tout comme la log-vraisemblance. En effet, certain phénomènes sont intrinsèquement complexes et même un bon modèle ne parviendra pas à rendre compte d'une grande partie de la variabilité de la réponse. Ce n'est pas non plus parce que le R^2 est faible que Y et et les variables explicatives X_j sont indépendantes, comme l'illustre la Figure 1.2.

En outre, il est possible de gonfler la valeur de R^2 en incluant davantage de variables explicatives et en rendant le modèle plus complexe, ce qui améliore la vraisemblance et R^2. En effet, le coefficient n'est pas décroissant dans la dimension de $\mathbf{X}$, de sorte qu'un modèle comportant $p+1$ de covariables aura nécessairement des valeurs de R^2 plus élevées que si l'on n'incluait que p de ces variables explicatives. Pour comparer les modèles, il est préférable d'utiliser des critères d'information ou de s'appuyer sur la performance prédictive si tel est l'objectif de la régression. Enfin, un modèle avec un R^2 élevé peut impliquer une corrélation élevée, mais la relation peut être fallacieuse: la régression linéaire ne produit pas de modèles causaux!

Bibliographie

Baumann, James F., Nancy Seifert-Kessell, et Leah A. Jones. 1992. « Effect of Think-Aloud Instruction on Elementary Students' Comprehension Monitoring Abilities ». *Journal of Reading Behavior* 24 (2): 143-72. https://doi.org/10.1080/10862969209547770.

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, et Sylvain Sénécal. 2021. « Smartwatches are more distracting than mobile phones while driving: Results from an experimental study ». *Accident Analysis & Prevention* 149: 105846. https://doi.org/10.1016/j.aap.2020.105846.

Brucks, Melanie S., et Jonathan Levav. 2022. « Virtual communication curbs creative idea generation ». *Nature* 605 (7908): 108-12. https://doi.org/10.1038/s41586-022-04643-y.

Davison, A. C. 2003. *Statistical Models*. Cambridge University Press.

Duke, Kristen E., et On Amir. 2023. « The Importance of Selling Formats: When Integrating Purchase and Quantity Decisions Increases Sales ». *Marketing Science* 42 (1): 87-109. https://doi.org/10.1287/mksc.2022.1364.

Gosset, William Sealy. 1908. « The probable error of a mean ». *Biometrika* 6 (1): 1-25. https://doi.org/10.1093/biomet/6.1.1.

Lee, Kiljae, et Jungsil Choi. 2019. « Image-text inconsistency effect on product evaluation in online retailing ». *Journal of Retailing and Consumer Services* 49: 279-88. https://doi.org/10.1016/j.jretconser.2019.03.015.

Liu, Peggy J., SoYon Rim, Lauren Min, et Kate E. Min. 2023. « The surprise of reaching out: Appreciated more than we think. » *Journal of Personality and Social Psychology* 124 (4): 754-71. https://doi.org/10.1037/pspi0000402.

McCullagh, P., et J. A. Nelder. 1989. *Generalized linear models*. Second edition. London: Chapman & Hall.

Moon, Alice, et Eric M VanEpps. 2023. « Giving Suggestions: Using Quantity Requests to Increase Donations ». *Journal of Consumer Research* 50 (1): 190-210. https://doi.org/10.1093/jcr/ucac047.

Rosen, B., et T. H. Jerdee. 1974. « Influence of sex role stereotypes on personnel decisions. » *Journal of Applied Psychology* 59: 9-14.

Sokolova, Tatiana, Aradhna Krishna, et Tim Döring. 2023. « Paper Meets Plastic: The Perceived Environmental Friendliness of Product Packaging ». *Journal of Consumer Research* 50 (3): 468-91. https://doi.org/10.1093/jcr/ucad008.

Venables, William N. 2000. « Exegeses on Linear Models ». In *S-PLUS User's Conference*. Washington, D.C. https://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf.