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 and 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.

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

- Si 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 and 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)?
- L'impact de de l'inadéquation entre l'image d'un produit et sa description (Lee and Choi 2019)
- Études supérieures: est-ce que le prix en vaut la chandelle?.
- 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? 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, and Döring 2023).
- L’impact psychologique des suggestions sur le montant de dons (Moon and VanEpps 2023)
- Si la visioconférence réduit le nombre d’interactions et d’idée créatives générées lors d’une réunion, par rapport à une rencontre en personne (Brucks and Levav 2022)

1 Introduction

1.1 Rappels

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

1.1.1 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 une collection 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).

1.1.2 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.

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.

Les modèles de régression servent à expliquer des variables quantitatives en fonction d'autres caractéristiques.

- 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.

1.1.3 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 fera la distinction entre deux cas de figure: quand le phénomène prend des valeurs finies et dénombrables, comme par exemple un événement binaire (achat/non-achat d'un produit) ou d'une variable de décompte, on parle de distribution **discrète**. Si l'ensemble des valeurs que prend la variable d'intérêt est dans un continuum de valeurs (par exemple, le prix d'un item, même si en pratique il y aura une sous-division finie et que le montant sera arrondi au centime près pour l'argent), on parle de variable **continue**. Des cas plus complexes peuvent survenir.

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

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

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

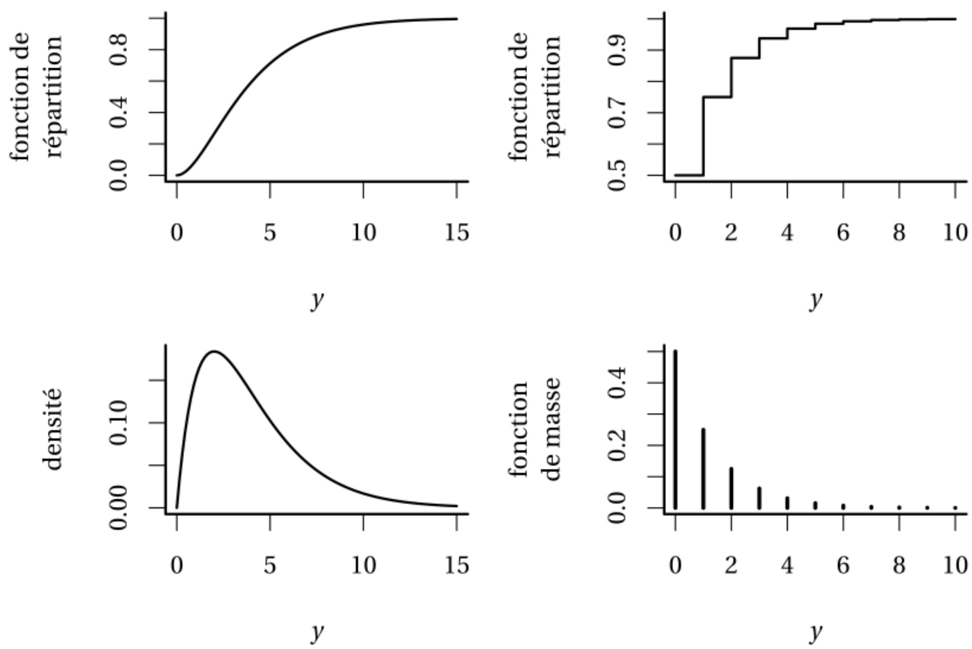

Figure 1.1: Fonctions de répartition (panneau supérieur) et fonctions de densité et de masse (panneau inférieur) pour une loi continue (gauche) et discrète (droite).

peuvent prendre Y. Plus généralement, l'espérance d'une fonction $g(y)$ pour une variable 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

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

Définition 1.3 (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!

Définition 1.4 (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. On peut chercher un estimateur qui minimise l'erreur quadratique moyenne,

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

C'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

1.1.4 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 1.5 (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 géométrique)**.** La loi géométrique décrit le comportement du nombre d'essais Bernoulli de probabilité de succès p nécessaires avant l'obtention d'un premier succès. La fonction de masse de $Y \sim \mathsf{Geo}(p)$ est

$$\mathsf{Pr}(Y = y) = p(1-p)^{y-1}, \quad y = 1, 2, \ldots$$

Par exemple, on pourrait modéliser le nombre de visites d'une maison en vente avant une première offre d'achat à l'aide d'une variable géométrique.

Exemple 1.5 (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 1.6 (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 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.

1.1.5 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 1.6 (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 $\alpha = \beta = 1$ correspond à la loi standard uniforme.

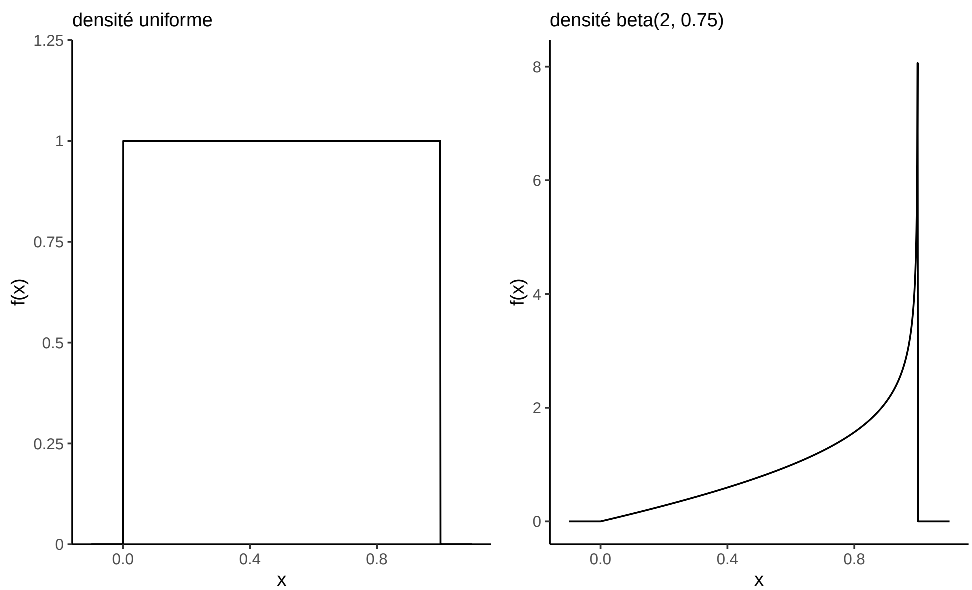

Figure 1.2: Fonctions de densité de lois uniformes et beta(2, 3/4) sur l'intervalle [0,1].

Définition 1.7 (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 1.8 (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 fonction de densité est symmétrique autour d'une moyenne $\mu \in \mathbb{R}$, qui est une mesure de la tendance centrale. L'autre paramètre, l'écart-type $\sigma > 0$, est un paramètre d'échelle qui mesure la dispersion. La fonction de densité, en forme de cloche, est symmétrique autour de μ, qui est aussi le mode de la distribution.

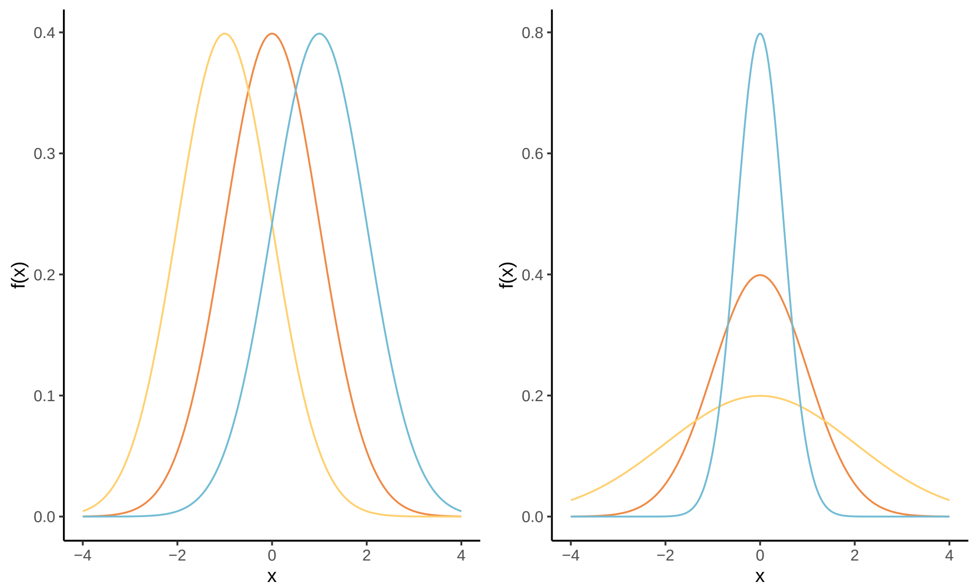

Figure 1.3: Densités de loi normales avec des paramètres de moyenne différents (gauche) et des paramètres d'échelle différents (droite).

La loi normale est une famille de localisation échelle: si $Y \sim \mathsf{normale}(\mu, \sigma^2)$, alors $Z = (Y - \mu)/\sigma \sim \mathsf{normale}(0, 1)$. La variable standardisée est appelée cote Z. Inversement, si $Z \sim \mathsf{normale}(0, 1)$, alors $Y = \mu + \sigma Z \sim \mathsf{normale}(\mu, \sigma^2)$.

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 1.9 (Loi khi-deux)**.** La loi de khi-deux avec $\nu > 0$ degrés de liberté, dénotée χ^2_ν joue un rôle important en statistique car elle est la loi nulle asymptotique de nombreuses statistiques de test usuelles. 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 en prenant la somme de variables normales centrées et réduites au carré: si $Z_i \stackrel{\text{iid}}{\sim} \mathsf{normale}(0,1)$ pour $i=1,\ldots,k$, alors $\sum_{i=1}^k Z_i^2 \sim \chi^2_k$. L'espérance de la loi χ^2_k est k.

Définition 1.10 (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. Ses ailes sont plus lourdes que la loi normale et son ke moment existe uniquement si $\nu > k$. Ainsi, la variance n'existe que si $\nu > 2$. Si $\nu \to \infty$, on recouvre une loi limite Gaussienne standard.

Son nom provient d'un article de William Gosset sous le pseudonyme Student (Gosset 1908). La statistique T, pour des données Gaussiennes, suit une loi Student avec k degrés de liberté, puisque la moyenne est Gaussienne et la variance empirique, adéquatement repondérée, suit une loi khi-deux avec k degrés de liberté.

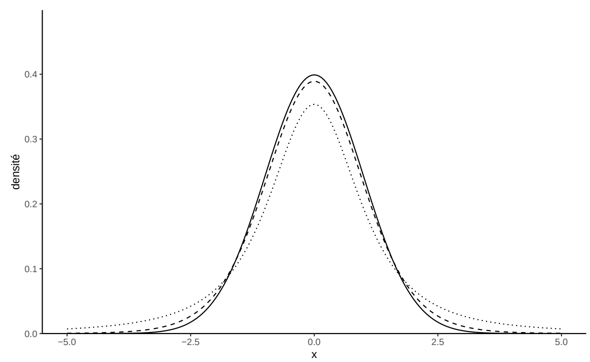

Figure 1.4: Comparaison de la densité Student-t versus normale pour différents degrés de liberté avec $\nu = 2$ (pointillé), $\nu = 10$ (traitillé) et la loi normale ($\nu = \infty$).

Définition 1.11 (Loi de Fisher)**.** Une généralisation de la loi Student-t est obtenue pour 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 χ^2 quand $\nu_2 \to \infty$.

1.1.6 Graphiques

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!

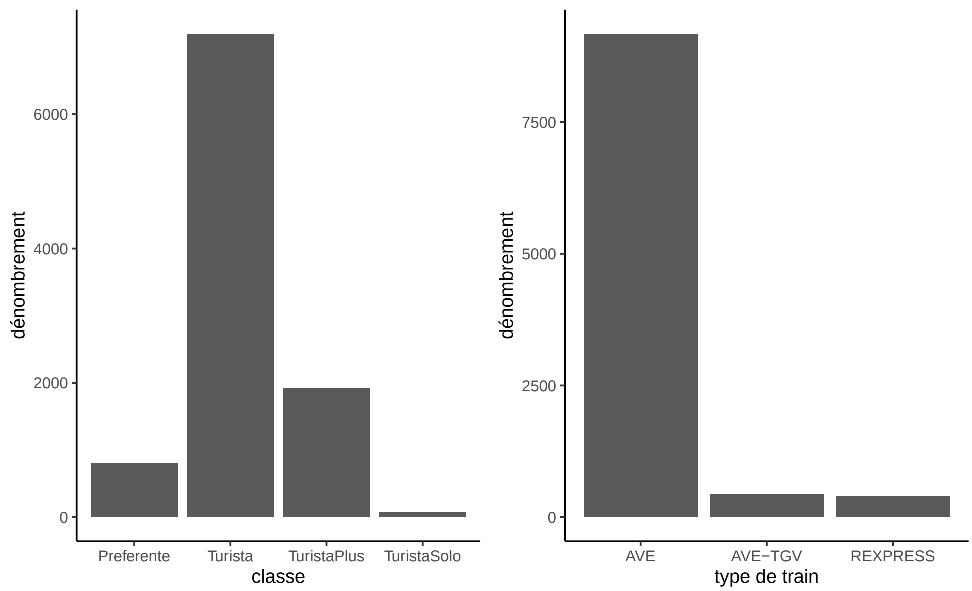

Figure 1.5: Diagramme en bâtons pour la classe des billets de trains du jeu de données Renfe.

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.

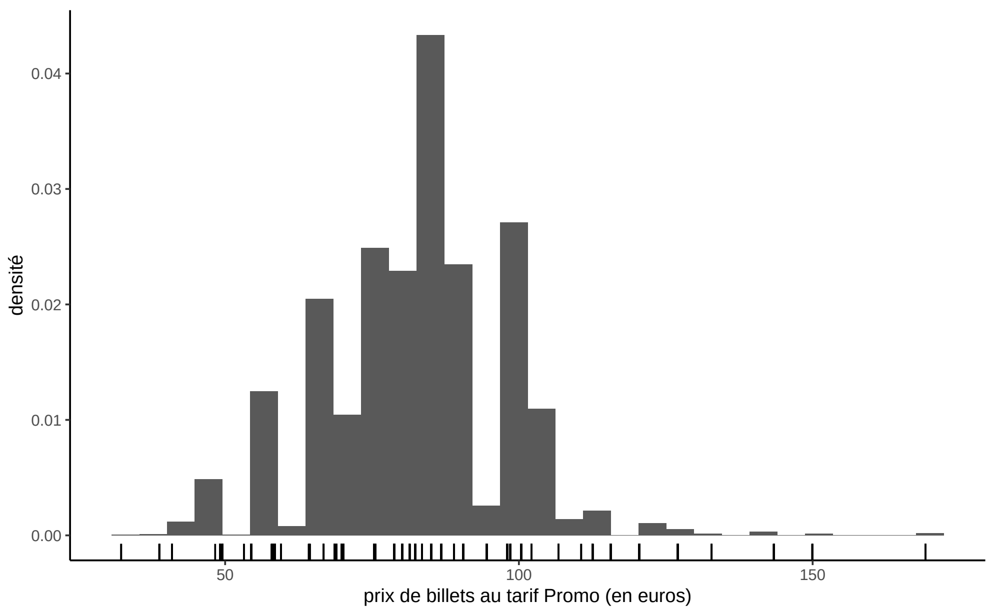

Figure 1.6: Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe

Définition 1.12 (Boîte à moustaches)**.** Elle représente graphiquement cinq statistiques descriptives.

- 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.

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

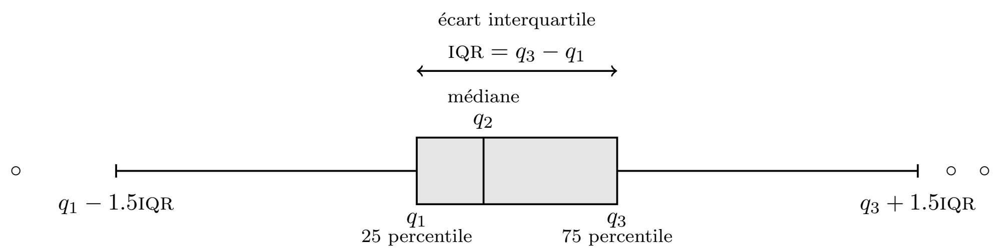

Figure 1.7: Boîte à moustache.

disposant côte-à-côte. Une troisième variable catégorielle peut être ajoutée par le biais de couleurs, comme dans la Figure 1.8.

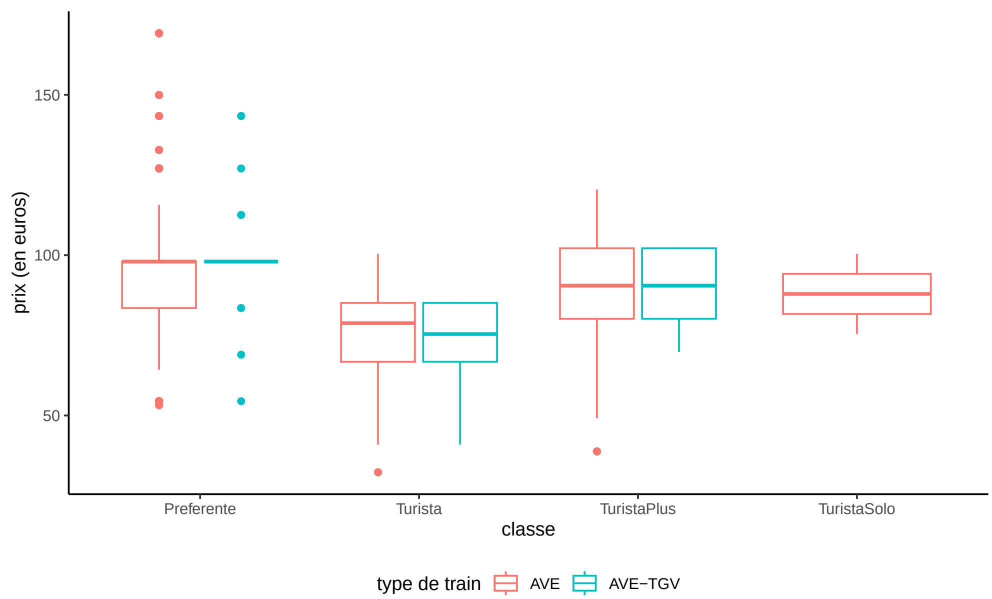

Figure 1.8: Boîte à moustaches du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe.

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.

Définition 1.13 (Diagrammes quantiles-quantiles)**.** 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. 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 $F(Y) \sim \mathsf{Unif}(0,1)$. De la même façon, appliquer la fonction quantile à une variable uniforme permet de simuler de la loi F, et donc $F^{-1}(U)$.

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 $\{\mathrm{rang}(Y_i)/(n+1), \widehat{F}(Y_i)\}$.

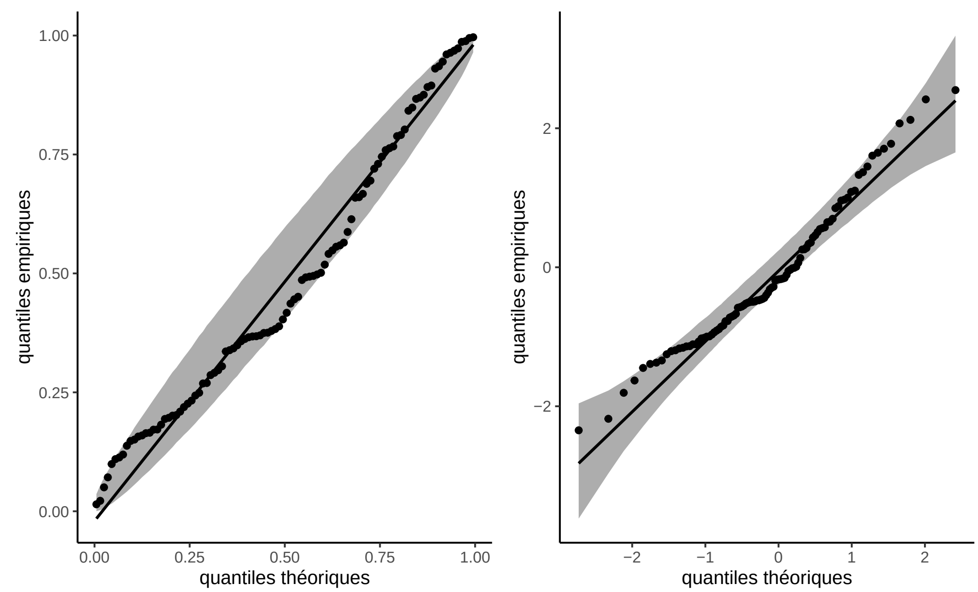

Figure 1.9: Diagramme probabilité-probabilité (gauche) et quantile-quantile normal (droite)

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 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 (autoamorçage paramétrique), 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 1.9 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.

1.1.7 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 $\hat{\theta} \stackrel{\mathsf{Pr}}{\to} \theta$: 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.

1.1.8 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 $\sqrt{n}(\overline{Y}-\mu) \stackrel{\mathrm{d}}{\rightarrow} \mathsf{normale}(0,\sigma^2)$. Si n est grand (typiquement supérieur à 30, mais cette règle dépend de la loi sous-jacente de Y), alors $\overline{Y} \stackrel{\cdot}{\sim} \mathsf{normale}(\mu, \sigma^2/n)$.

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.

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 1.10 montre la distribution empirique des données.

Considérons maintenant des échantillons de taille $n=10$. Dans notre premier échantillon aléatoire, la durée moyenne affichée est 170.9 minutes, elle est de 164.5 minutes dans le deuxième, de 172.3 dans le troisième, et ainsi de suite.

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 1.11 montre un de ces 1000 échantillons aléatoire de taille $n=20$ tiré de notre base de données. Les autres graphiques de la Figure 1.11 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

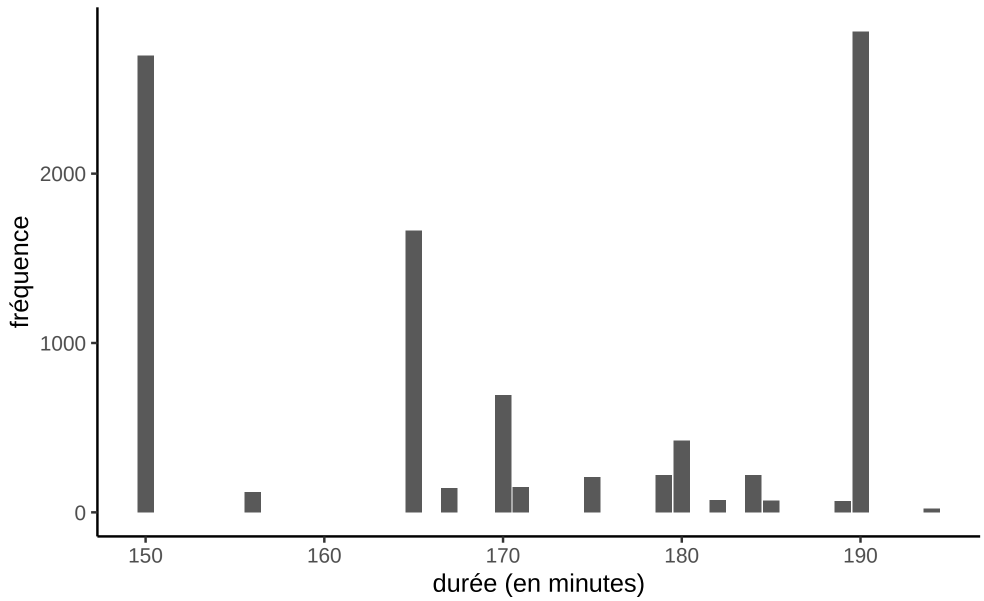

Figure 1.10: Distribution empirique des temps de trajet en trains à grande vitesse.

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!

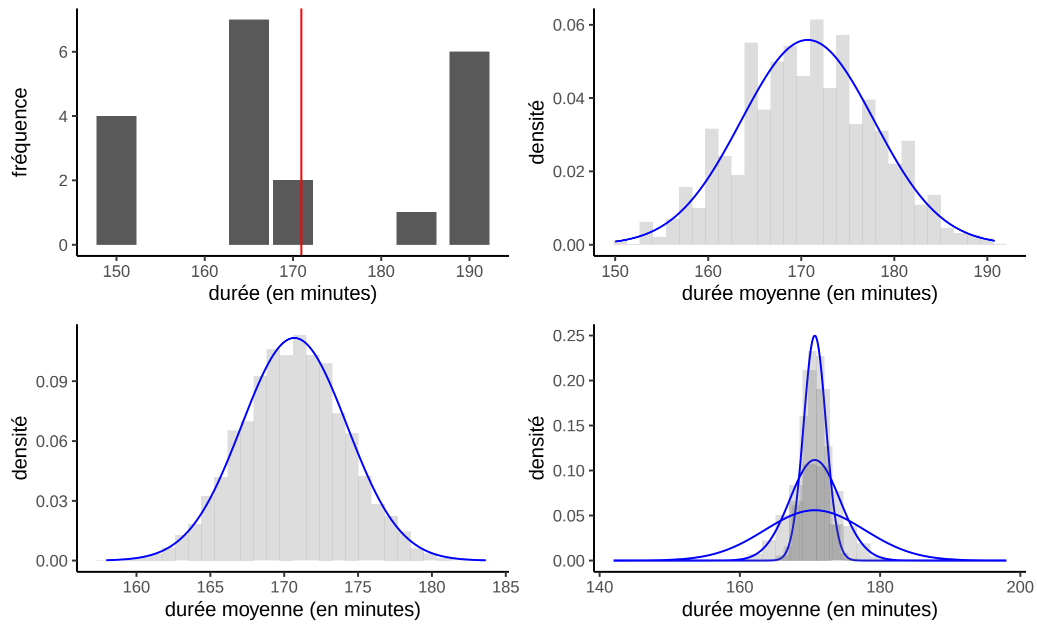

Figure 1.11: Représentation graphique du théorème central limite: échantillon aléatoire de 20 observations avec leur moyenne empirique (trait vertical rouge) (en haut à gauche). Les trois autres panneaux montrent les histogrammes des moyennes empiriques d'échantillons répétés de taille 5 (en haut à droite), 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.

2 Inférence statistique

2.0.1 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. Le matériel de la section suivante a été préparé par Juliana Schulz.

2.0.2 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?

Remarque: 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 $\mathscr{H}_a : \theta < \theta_0$.

2.0.3 Statistique de test

Une statistique de test T est un fonctionnel 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

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

Il convient de faire la différence entre procédures/formules et valeurs numériques. Un **estimateur** est une règle ou une formule utilisée pour calculer l'estimation d'un paramètre ou quantité d'intérêt selon des données observées. Par exemple, la moyenne d'échantillon $\overline{X}$ est un estimateur de la moyenne dans la population μ. Une fois qu'on a des données observées, on peut calculer un estimé de la moyenne empirique $\overline{x}$, c'est-à-dire, on obtient une valeur numérique. Autrement dit,

- un estimateur est une fonction de variables aléatoires et donc c'est aussi une variable aléatoire car sa valeur fluctue d'un échantillon à l'autre.
- l'estimé est la valeur numérique calculée sur un échantillon donné.

2.0.4 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 $\mathsf{Pr}_0(|T| \geq |t|)$. Si la distribution de T est symétrique autour de zéro, la valeur-*p* vaut

$$p = 2 \times \mathsf{Pr}_0(T \geq |t|).$$

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 $\mathsf{Norm}(\mu, \sigma^2)$, 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 St_{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 $P = 2 \times \mathsf{Pr}(T > |t|)$, où $T \sim \mathsf{St}_{n-1}$.

2.0.5 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.

$$\alpha = \mathsf{Pr}_0\left(\text{ rejeter } \mathscr{H}_0\right).$$

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.

2.0.6 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

$$\gamma = \mathsf{Pr}_a(\text{rejeter } \mathscr{H}_0)$$

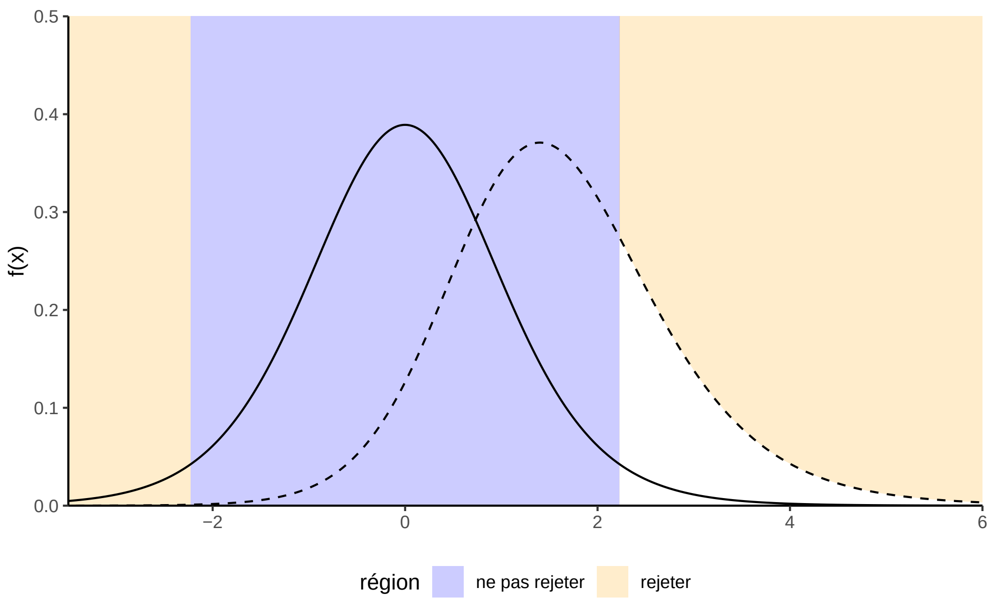

Figure 2.1: Comparaison de la loi nulle (ligne pleine) et d'une alternative spécifique pour un test-t (ligne traitillée). La puissance correspond à l'aire sous la courbe de la densité de la loi alternative qui est dans la zone de rejet du test (en blanc).

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, on veut que γ soit 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 du paramètre postulé θ_0 sous $\mathscr{H}_0$ et le comportement observé, plus il est facile de le détecter (voir Figure 2.3);
- 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 la Figure 2.2);
- 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

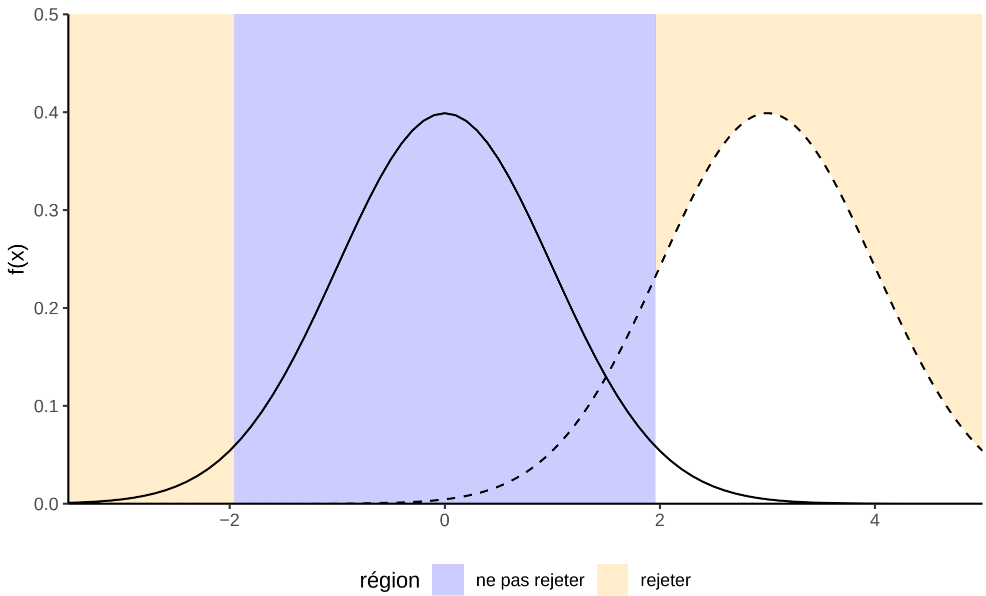

Figure 2.2: Augmentation de la puissance suite à une augmentation de la différence de moyenne sous l'hypothèse alternative. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée); cette dernière est plus décalée vers la droite par rapport à la loi nulle postulée (ligne pleine).

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 pour la puissance: par exemple, si on utilise un test-t pour un échantillon, la statistique $T = \sqrt{n}(\overline{X} - \mu_0)/S_n \sim \mathcal{T}_{n-1}$ et, si la vraie moyenne est $\Delta + \mu_0$, alors la loi alternative est Student-t, mais non-centrée avec paramètre de décalage Δ. Cette dérivation est l'exception plutôt que la règle et on détermine d'ordinaire 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.

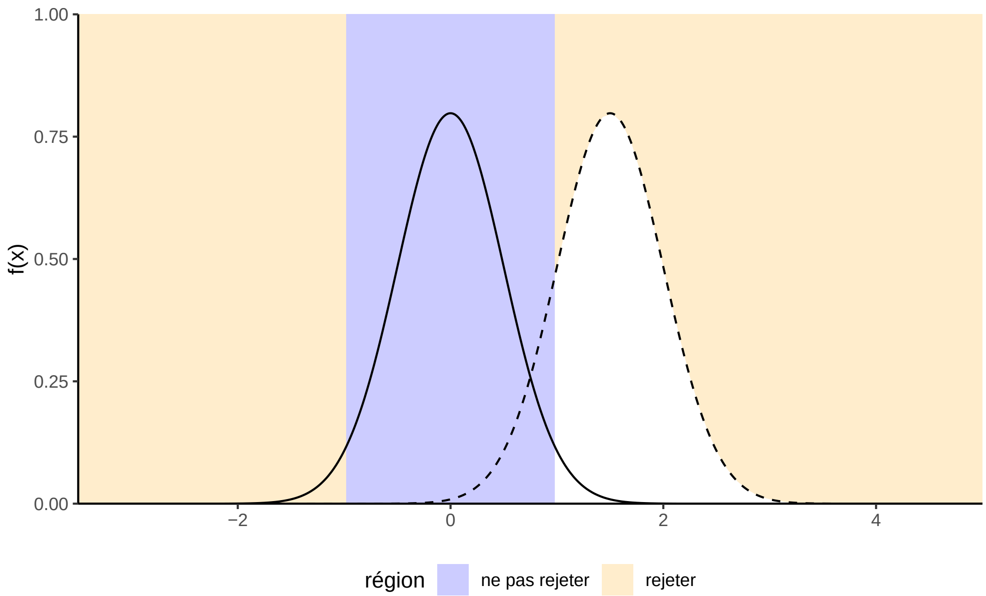

Figure 2.3: Augmentation de la puissance suite à une augmentation de la taille de l'échantillon ou une diminution de l'écart-type de la population: la loi nulle (ligne pleine) est plus concentrée et la taille de la région de rejet diminue. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée). Règle générale, la loi nulle change selon la taille de l'échantillon.

2.0.7 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

$$\widehat{\theta} \pm \mathfrak{q}_{\alpha/2}\, \mathrm{se}(\widehat{\theta})$$

où $\mathfrak{q}_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ 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. 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.

Par exemple, pour un échantillon aléatoire $X_1, \ldots, X_n$ provenant d'une loi normale $\mathsf{Norm}(\mu, \sigma)$, 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.

Avant qu'on calcule l'intervalle de confiance, il y a une probabilité de $1-\alpha$ que θ soit contenu dans l'intervalle **aléatoire** symétrique $(\widehat{\theta} - \mathfrak{q}_{\alpha/2}\,\mathrm{se}(\widehat{\theta}), \widehat{\theta} + \mathfrak{q}_{\alpha/2}\,\mathrm{se}(\widehat{\theta}))$, 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.

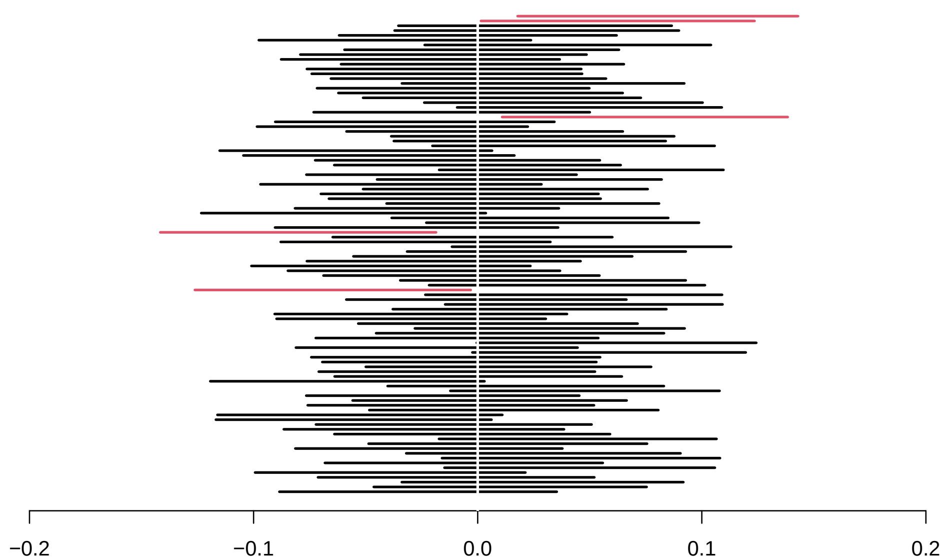

Figure 2.4: Intervalles de confiance à 95% pour la moyenne d'une population normale $\mathsf{Norm}(0,1)$ pour 100 échantillons aléatoires. En moyenne, 5% de ces intervalles (en rouge) n'incluent pas la vraie valeur de la moyenne de zéro.

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 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.

Exemple 2.1 (Achat en ligne de milléniaux)**.** Supposons qu'une chercheuse veut faire une étude sur l'évolution des ventes en ligne au Canada. Elle postule que les membres de la génération Y fait plus d'achats en ligne que ceux des générations antérieures. Pour répondre à cette question, un sondage est envoyé à un échantillon aléatoire de $n=500$ individus représentatif de la population avec 160 membres de la génération Y et 340 personnes plus âgées. La variable réponse est le montant d'achat effectués en ligne dans le mois dernier (en dollars).

Dans cet exemple, on s'intéresse à la différence entre le montant moyen des Y et celui des générations antérieures: la différence de moyenne observée dans l'échantillon est de 16.49 dollars et donc les milléniaux ont dépensé davantage. En revanche, notre échantillon est aléatoire et le montant d'achat en ligne varie d'un individu à l'autre (et d'un mois à l'autre): ce n'est donc pas suffisant pour dire que la différence est significative.

La première étape de notre analyse consiste à définir les quantités d'intérêt et à formuler nos hypothèse en fonction de paramètres du modèle; il convient également de définir ces derniers en fonction des variables en présence dans l'exemple. Ici, on considère un test pour la différence de moyenne dans les populations postulées μ_1 (pour la génération Y) et μ_2 (pour les générations antérieures) d'écart-type respectif σ_1 et σ_2. Comment déterminer quelle hypothèse on considère? Comme statisticien, on se fait l'avocat du Diable: l'hypothèse d'intérêt du chercheur est l'hypothèse alternative et ici, $\mathscr{H}_a : \mu_1 > \mu_2$, où μ_1 représente la moyenne des achats mensuels des milléniaux. L'hypothèse nulle comprend toutes les autres valeurs pour la différence de moyenne, soit $\mathscr{H}_0 : \mu_1 \leq \mu_2$. Il suffit néanmoins de considérer le cas $\mu_1 = \mu_2$ (pourquoi?)

La deuxième étape consiste à choisir une statistique de test. S'il n'y a aucune différence de moyenne entre les groupes, alors $\overline{X}_1 - \overline{X}_2$ a moyenne zéro et la différence de moyenne a une variance de $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Ici, on considère la statistique de Welch (1947) pour une différence de moyenne entre deux échantillons:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{1/2}},$$

où $\overline{X}_i$ est la moyenne empirique dans l'échantillon i ($i = 1, 2$) et S_i^2 est la variance empirique et n_i la taille de l'échantillon du groupe i. La statistique est utilisée pour calculer la différence de moyennes de deux échantillons de variance potentiellement différente. La valeur de la statistique dans l'échantillon est $T = 2.76$, mais on obtiendrait une valeur différente avec un autre échantillon. Il convient donc de déterminer si cette valeur est compatible avec notre hypothèse nulle en la comparant à la loi nulle sous $\mathscr{H}_0$ de T. On effectuera le test à niveau $\alpha = 0.05$.

La troisième étape est l'obtention d'un étalon de mesure pour déterminer si notre résultat est extrême ou inattendu. Vous remarquerez que la statistique de Welch a moyenne zéro et variance un sous l'hypothèse nulle que $\mu_1 = \mu_2$: standardiser une statistique permet d'obtenir un objet dont on connaît le comportement pour de grands échantillons et obtenir une quantité sans unité de mesure. La dérivation de la loi nulle est hors objectifs du cours, aussi cette dernière vous sera donnée dans tous les cas qu'on considère. Asymptotiquement, T suit une loi normale $\mathsf{Norm}(0, 1)$, mais il existe une meilleure approximation pour n petit; on compare le comportement de T à l'aide d'une loi de Student via l'approximation de Satterthwaite (1946).

La dernière étape consiste à obtenir une valeur-*p*, soit la probabilité d'observer un résultat aussi extrême sous $\mathscr{H}_0$: l'avantage de la valeur-*p* est que cette valeur est une probabilité (dans $[0, 1]$) et qu'elle suit une loi uniforme sous $\mathscr{H}_0$. Puisque nous avons une hypothèse alternative unilatérale, on regarde la probabilité sous $\mathscr{H}_0$ que $\mathsf{Pr}(T > t)$. La valeur-*p* vaut 0.0031 et donc, à niveau 5%, on rejette l'hypothèse nulle pour conclure que la génération Y dépense davantage en ligne que les générations antérieures.

Exemple 2.2 (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.

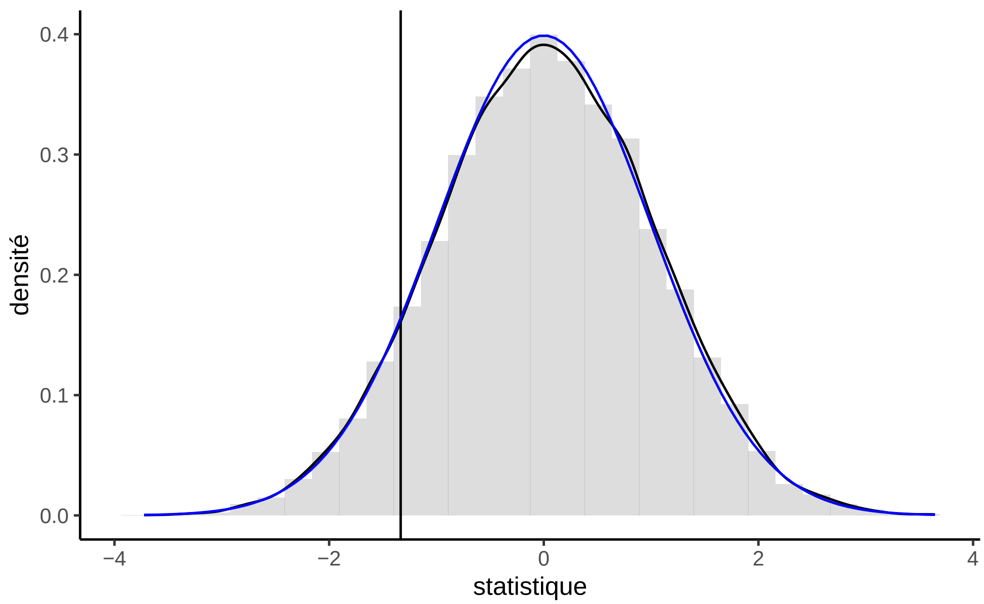

Figure 2.5: Approximation par permutation de la loi nulle de la statistique de test de Welch (histogramme et trait noir) et loi asymptotique normale standard (trait bleu) pour le prix de billets de trains AVE au tarif promotionnel entre Madrid et Barcelone. La valeur de la statistique de test de l'échantillon original est représentée par un trait vertical.

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 2.5. Malgré que notre échantillon soit très grand, avec $n = 8059$ observations, la différence n'est pas 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.

Bibliographie

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, and 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., and Jonathan Levav. 2022. "Virtual Communication Curbs Creative Idea Generation." *Nature* 605 (7908): 108–12. https://doi.org/10.1038/s41586-022-04643-y.

Duke, Kristen E., and 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, and 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.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. Second edition. London: Chapman & Hall.

Moon, Alice, and 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.

Satterthwaite, F. E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2 (6): 110–14. http://www.jstor.org/stable/3002019.

Sokolova, Tatiana, Aradhna Krishna, and 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.

Welch, B. L. 1947. "The Generalization of 'Student's' Problem When Several Population Variances Are Involved." *Biometrika* 34 (1–2): 28–35. https://doi.org/10.1093/biomet/34.1-2.28.