

Modélisation statistique

Table des matières

Remarques	5
1 Introduction à l'inférence statistique	7
1.1 Tests d'hypothèse	8
1.2 Analyse exploratoire de données	20
2 Régression linéaire	35
3 Modèles linéaires généralisés	37
4 Données corrélées et longitudinales	39
5 Modèles linéaires mixtes	41
6 Analyse de survie	43
7 Inférence basée sur la vraisemblance	45
R	47
A Compléments mathématiques	49
A.1 Variables aléatoires	50
A.2 Loi des grands nombres	53
A.3 Théorème central limite	53

Remarques

Ces notes sont l'oeuvre de Léo Belzile (HEC Montréal) et sont mis à disposition sous la Licence publique Creative Commons Attribution - Utilisation non commerciale - Partage dans les mêmes conditions 4.0 International et ont été compilé le 23 juillet 2020.

Bien que les diapositives illustrent l'implémentation des techniques statistiques et des modèles à l'aide de **SAS**, ces notes présentent le pendant **R** : visitez le site web du projet **R** pour télécharger le logiciel. L'interface graphique la plus populaire (et celle que je vous recommande) est RStudio Desktop.

Ce cours traite de modélisation des données, aussi convient-il de s'attarder au fait que nos modèles seront des approximations de la réalité : George Box a affirmé à plusieurs reprises que « tous les modèles sont faux, mais certains sont utiles ». Ce point de vue est réducteur; Peter McCullagh et John Nelder (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; certains, cependant, sont mieux que d'autres 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.

Chapitre 1

Introduction à l'inférence statistique

Ce chapitre porte sur deux concepts fondamentaux pour la modélisation, à savoir les principes sous-jacents aux tests d'hypothèses et l'analyse exploratoire des données.

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 pour valider ou infirmer cette dernière et conclure 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 déjà été collectées, on parle de données **observationnelles**.

Il y a généralement deux raisons de développer un modèle reliant une variable réponse Y à un ensemble de variables explicatives \mathbf{X} : à des fins de prédiction (modèle prédictif) ou pour tester des hypothèses de recherche concernant les effets de ces variables (modèle explicatif). On fera dans ce cours une distinction entre inférence et prédiction, bien que ces deux objectifs ne soient pas mutuellement exclusifs. 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 ; ainsi, les modèles de régressions sont fréquemment utilisés pour l'inférence.

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. On se concentrera dans ce cours sur les modèles explicatifs. Par exemple, on peut chercher à déterminer

- Est-ce que les consommateurs sont prêts à dépenser davantage lorsqu'ils paient par crédit qu'en argent comptant?
- Est-ce qu'il y a de la discrimination salariale envers les femmes professeurs d'un collège américain?
- Études supérieures : est-ce que le prix en vaut la chandelle?.
- Quels sont les critères médicaux qui impactent les primes d'assurance maladies?
- 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 sont plus faciles en régions en Grande-Bretagne? Une analyse du journal britannique *The Guardian* laisse penser que c'est le cas.
- Est-ce le risque de transmission de la Covid augmente en fonction de la distanciation? Une (mauvaise) méta-analyse dit que oui (ou l'art de tirer des conclusions erronées à partir d'une étude bancale).

1.1 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 hypothèses que l'on veut tester en fonction de paramètres du modèle,
- calculer la statistique de test,
- déterminer son comportement sous \mathcal{H}_0 (loi nulle),
- calculer la valeur- p ,
- conclure dans le contexte du problème (rejeter ou ne pas rejeter \mathcal{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 \mathcal{H}_0 est donc *non-coupable*, et l'hypothèse alternative \mathcal{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 maxi-

miser 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 \mathcal{H}_0 : si la personne était innocente, est-ce que les preuves présentées tiendraient la route ? Des preuves probantes (ADN, etc.) auront davantage de poids que des preuves circonstancielles (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 \mathcal{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.

1.1.1 Hypothèse

Dans les test statistique il y a toujours deux hypothèse : l'hypothèse nulle (\mathcal{H}_0) et l'hypothèse alternative (\mathcal{H}_a). Habituellement, l'hypothèse nulle est le « statu quo » et l'alternative est l'hypothèse que l'on cherche à démontrer. Un test d'hypothèse statistique nous permet de décider si nos données nous fournissent assez de preuves pour rejeter \mathcal{H}_0 en faveur de \mathcal{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 unidimensionnel θ s'exprimerait la forme suivante :

$$\mathcal{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \mathcal{H}_a : \theta \neq \theta_0.$$

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 dans la population, les hypothèses sont :

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

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 ?

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

1.1.2 Statistique de test

Une statistique de test T est une fonction des données d'échantillon qui contient de 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 \mathcal{H}_0 , c'est-à-dire l'ensemble des valeurs que prend T si \mathcal{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 \mathcal{H}_0 est vraie. Plusieurs statistiques que l'on couvrira dans ce cours sont des **statistiques de Wald**, de la forme

$$T = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

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

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

$$\mathcal{H}_0 : \mu = 0 \quad \text{versus} \quad \mathcal{H}_a : \mu \neq 0,$$

la statistique de test de Wald est

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

où \bar{X} est la moyenne de l'échantillon X_1, \dots, X_n ,

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

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

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

Il convient de faire la différence entre fonctions 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 \bar{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 \bar{x} , c'est-à-dire, on obtient une valeur numérique. Autrement dit,

- un estimateur est la procédure ou formule qui nous dit comment utiliser les données pour calculer une estimation.
- un estimateur est une variable aléatoire car sa valeur fluctue d'un échantillon à l'autre.
- l'estimé est la valeur numérique calculée sur un échantillon donné.

1.1.3 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 \mathcal{H}_0 . Plus précisément, la valeur- p est la probabilité, si \mathcal{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, \dots, X_n et qu'on observe une valeur de la statistique de test de $T = t$. Pour un test d'hypothèse bilatéral $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_a : \theta \neq \theta_0$, la valeur- p est $\Pr_0(|T| \geq |t|)$, c'est-à-dire, la probabilité que $|T|$ est égal ou plus grand que ce qu'on observe, en valeur absolue, sous \mathcal{H}_0 . Si la distribution de T est symétrique autour de 0, la valeur- p vaut

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

Prenons l'exemple d'un test d'hypothèse bilatéral pour la moyenne au population $\mathcal{H}_0 : \mu = 0$ contre $\mathcal{H}_a : \mu \neq 0$. Si l'échantillon provient d'une (population de) loi normale $\text{No}(\mu, \sigma^2)$, on peut démontrer que, si \mathcal{H}_0 est vraie et donc $\mu = 0$, la statistique de test

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

suit une loi de Student- t avec $n - 1$ degrés de liberté. 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 $P = 2 \times \Pr(T_{n-1} > |t|)$, où T_{n-1} dénote une variable aléatoire avec distribution de Student- t avec $n - 1$ degrés de liberté.

1.1.4 Conclusion

La valeur- p nous permet de faire une décision quant aux hypothèses du test. Si \mathcal{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 est peu probable, et donc nous aurons tendance de croire que \mathcal{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 \mathcal{H}_0 alors que \mathcal{H}_0 est vraie
- erreur de type II : on ne rejette pas \mathcal{H}_0 alors que \mathcal{H}_0 est fausse

Si le modèle générant les données est correct (sic), alors l'hypothèse nulle ou l'hypothèse alternative est vraie (ces deux scénarios couvrant l'univers des possibles). Si on fait un test d'hypothèse, il y a toujours une possibilité de commettre une erreur judiciaire (dit erreur de Type 1, c'est-à-dire de condamner un innocent ou rejeter \mathcal{H}_0 alors que l'hypothèse nulle est vraie). Pour se prémunir de 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. Le choix du statu quo (typiquement \mathcal{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 la chloroquine, un antipaludique, soit efficace face au virus de la Covid19).

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

Comme chercheur, on doit fixer préalablement le niveau de risque que nous sommes prêt à tolérer. Si on connaît la distribution de T sous \mathcal{H}_0 , on peut contrôler le risque de faire un erreur de type I. Ceci fait référence au **niveau** du test, dénoté par α :

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

La valeur de $\alpha \in (0, 1)$ est la probabilité qu'on rejette \mathcal{H}_0 quand \mathcal{H}_0 est en fait vraie. Comme chercheur, on choisit ce niveau α ; habituellement 1%, 5% ou 10%. Pour prendre une décision, on doit comparer la valeur- p P avec le niveau du test α :

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

1.1.5 Puissance statistique

Le but du test d'hypothèse est de découvrir des différences ou des effets significatifs : notre hypothèse d'intérêt est typiquement représenté par l'hypothèse alternative et on fait tout notre possible pour arriver à détecter quand cette dernière est plausible : par exemple, si une nouvelle configuration d'un site web (hypothèse alternative) permet d'augmenter les ventes par rapport au statu quo (hypothèse nulle). Notre capacité à détecter cette amélioration dépend de la puissance du test : plus cette dernière est élevée, plus grande est notre capacité à rejeter \mathcal{H}_0 . Quand on ne rejette pas \mathcal{H}_0 et que \mathcal{H}_a est en fait vraie, on commet une erreur de type II. Dénotons par $1 - \gamma$ la probabilité de faire une erreur de type II, c'est-à-dire

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

La **puissance statistique** d'un test est la probabilité que le test rejette \mathcal{H}_0 alors que \mathcal{H}_0 est fausse, soit γ . Selon le choix de l'alternative, il est plus ou moins facile de rejeter l'hypothèse nulle en faveur de l'alternative.

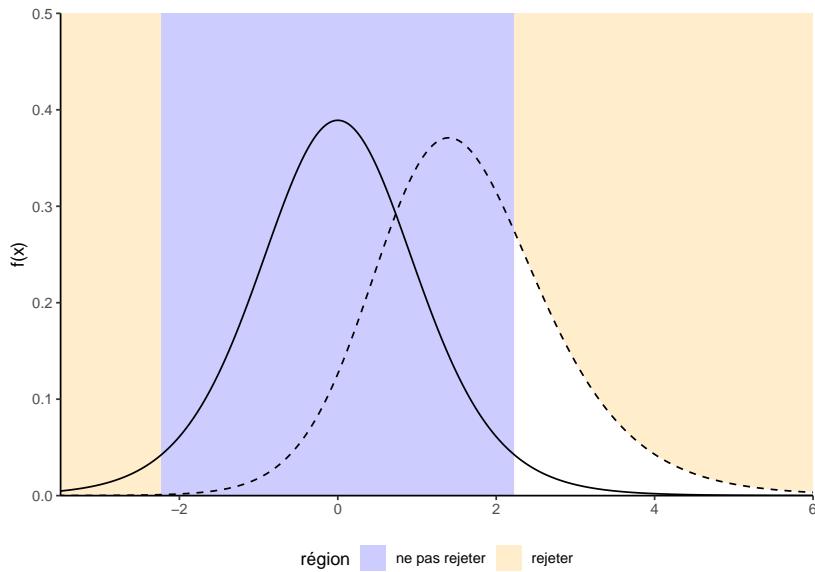


FIGURE 1.1 – Comparaison de la loi nulle (ligne pleine) et d'une alternative spécifique pour un test t (ligne traitillée). La puissance correspond à l'aire sous la courbe de la densité de la loi alternative qui est dans la zone de rejet du test (en blanc).

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 parce qu'on rejette l'hypothèse nulle $\alpha\%$ du temps même quand cette dernière est vraie. Aux fins de démonstration, il faut amasser suffisamment de preuves : la puissance, qui correspond à notre habileté à détecter quand \mathcal{H}_0 est fausse, 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 \mathcal{H}_0 et le comportement observé, plus il est facile de le détecter (voir Figure 1.3) ;
- 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 1.2) ;
- 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 influe aussi sur la puissance, mais 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 dans le cadre de ce cours.

Pour calculer la puissance d'un test, il faut choisir une alternative spécifique : par exemple, si on

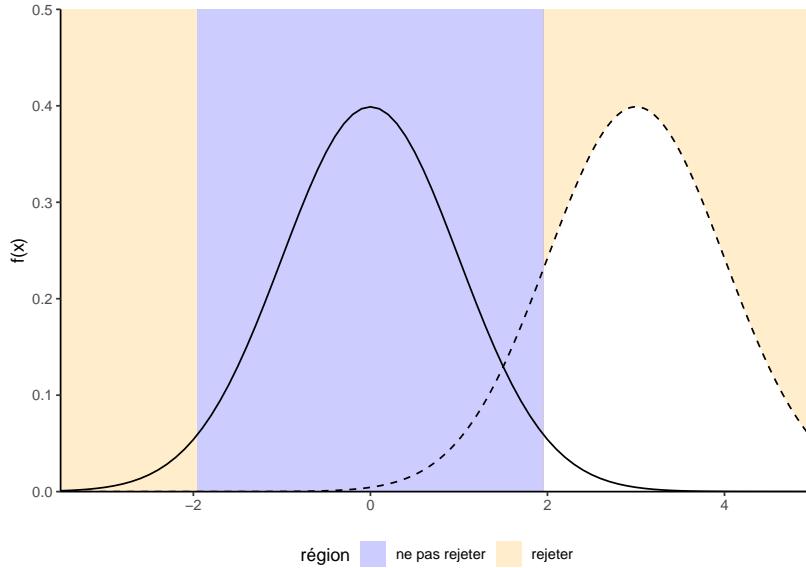


FIGURE 1.2 – Augmentation de la puissance suite à une augmentation de la différence de moyenne sous l'hypothèse alternative. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée); cette dernière est plus décalée vers la droite par rapport à la loi nulle postulée (ligne pleine).

utilise un test- t pour un échantillon, la statistique $T = \sqrt{n}(\bar{X} - \mu_0)/S_n \sim \mathcal{T}_{n-1}$. Si la vraie moyenne est $\Delta + \mu_0$, alors la loi alternative est Student- t , mais non-centrée avec paramètre de décalage Δ . Règle générale, on détermine la puissance à l'aide d'un estimateur Monte-Carlo en simulant des observations d'une alternative donnée, en calculant la statistique de test et la valeur- p associée 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 .

1.1.6 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

$$\hat{\theta} \pm q_{\alpha/2} \text{ se}(\hat{\theta})$$

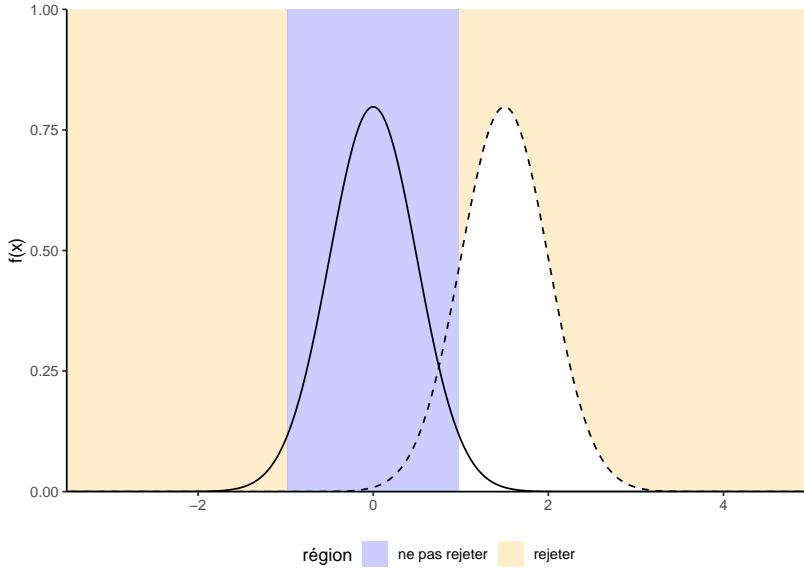


FIGURE 1.3 – Augmentation de la puissance suite à une augmentation de la taille de l'échantillon ou une diminution de l'écart-type de la population : la loi nulle (ligne pleine) est plus concentrée et la taille de la région de rejet diminue. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée). Règle générale, la loi nulle change selon la taille de l'échantillon.

où $q_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi nulle de la statistique de Wald T , soit

$$T = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})}$$

et θ représente la valeur du paramètre θ (supposé fixe, mais inconnu) de la population. Les bornes de l'intervalle de confiance sont aléatoires puisque $\hat{\theta}$ et $\text{se}(\hat{\theta})$ sont des variables aléatoires : leurs valeurs observées changent d'un échantillon à un autre.

Par exemple, pour un échantillon aléatoire X_1, \dots, X_n provenant d'une loi normale $\text{No}(\mu, \sigma)$, l'intervalle de confiance à $(1 - \alpha)$ pour la moyenne (dans la population) μ est

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

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 $(\hat{\theta} - q_{\alpha/2} \text{se}(\hat{\theta}), \hat{\theta} + q_{\alpha/2} \text{se}(\hat{\theta}))$. Une fois qu'on a 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 $(1 - \alpha)\%$ 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 grossièrement une fréquence de 50% de pile et 50% de face, mais chaque lancer donnera un ou l'autre de ces choix).

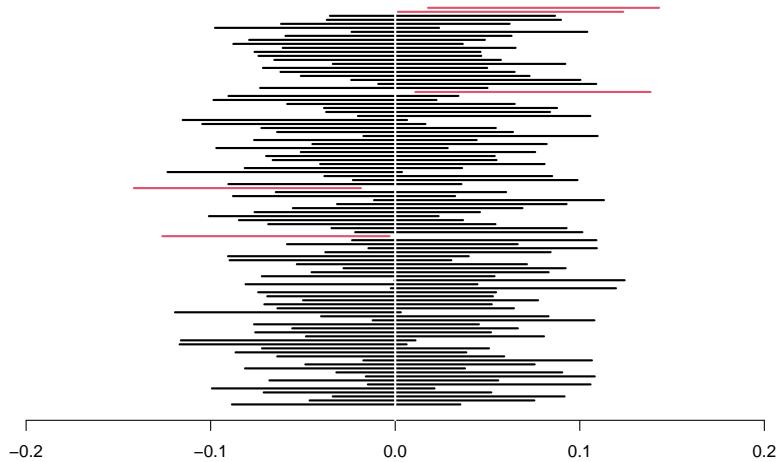


FIGURE 1.4 – Intervalles de confiance à 95% pour la moyenne d'une population normale $N(0, 1)$ pour 100 échantillons aléatoires. En moyenne, 5% de ces intervalles (en rouge) n'incluent pas la vraie valeur de la moyenne de zéro.

Si on s'intéresse seulement à la décision rejeter/ne pas rejeter \mathcal{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 \mathcal{H}_0 : pour un test à niveau α , on ne rejette pas \mathcal{H}_0 si et seulement si la valeur postulée pour θ est 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.

1.1.7 Exemple : achat en ligne de milléniaux

Supposons qu'un chercheur veut faire une étude sur l'évolution des ventes en ligne au Canada. Elle postule que les membres de la génération Y font 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, $\mathcal{H}_a : \mu_1 > \mu_2$, 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 $\mathcal{H}_0 : \mu_1 \leq \mu_2$. 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 $\bar{X}_1 - \bar{X}_2$ 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 :

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

où \bar{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 \mathcal{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 $\text{No}(0, 1)$, mais il existe une meilleure approximation pour n petit ; on compare le comportement de T à l'aide d'une loi de Student (à l'aide de l'approximation de Satterthwaite (1946)).

La dernière étape consiste à obtenir une valeur- p , soit la probabilité d'observer un résultat aussi extrême sous \mathcal{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 \mathcal{H}_0 . Puisque nous avons une hypothèse alternative unilatérale, on regarde la probabilité sous \mathcal{H}_0 que $\Pr(T > t)$. La p -valeur 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.

1.1.8 Exemple : prix de billets de trains

f 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 agrégés 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 $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$. Un test statistique nous permettrait de vérifier cette affirmation.

```
# Bibliothèque graphique
library(ggplot2)
# Manipulation de données, incluant %>%
library(poorman)
# Charger les données format SAS
url <- "https://lbelzile.bitbucket.io/MATH60604/renfe.sas7bdat"
renfe <- haven::read_sas(url)
head(renfe, n = 5)

## # A tibble: 5 x 7
##   prix type    classe     tarif      dest duree  jour
##   <dbl> <chr>   <chr>     <chr>     <dbl> <dbl>   <dbl>
## 1 143.  AVE    Preferente Promo      0    190     6
## 2 182.  AVE    Preferente Flexible  0    190     2
## 3 86.8   AVE    Preferente Promo      0    165     7
## 4 86.8   AVE    Preferente Promo      0    190     7
## 5 69.0   AVE-TGV Preferente Promo      0    175     4
```

```
#Identifier les variables catégorielles
renfe <- renfe %>%
```

```

mutate(dest = factor(recode(dest, "0" = "Barcelone-Madrid",
                            "1" = "Madrid-Barcelone")),
       type = factor(type),
       dest = factor(dest),
       jour = factor(jour),
       tarif = factor(tarif),
       classe = factor(classe))
# Sous-échantillon avec uniquement les données au tarif promotionnel
renfe_promo <- renfe %>% subset(tarif == "Promo")
# Test-t et différence de moyenne
ttest <- t.test(prix~dest, data = renfe_promo)
ttest #imprimer le résultat

```

```

##
## Welch Two Sample t-test
##
## data: prix by dest
## t = -1, df = 8040, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.100 0.209
## sample estimates:
## mean in group Barcelone-Madrid mean in group Madrid-Barcelone
##                               82.1                           82.6

```

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 nouvelles 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 10000). 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.

```

# Valeur-p par permutation
n <- nrow(renfe_promo)
B <- 1e4
ttest_stats <- numeric(B)
ttest_stats[1] <- ttest$statistic
set.seed(20200608) # germe pour nombres pseudo-aléatoires
for(i in 2:B){
  # Recalculer la statistique de test, mais permuter les étiquettes
  ttest_stats[i] <- t.test(prix ~ dest[sample.int(n = n)],
                           data = renfe_promo)$statistic
}
# Tracer un graphique de la distribution empirique obtenue par permutation
ggplot(data = data.frame(statistique = ttest_stats),
        aes(x=statistique)) +
  geom_histogram(bins = 30, aes(y=..density..), alpha = 0.2) +
  geom_density() +
  geom_vline(xintercept = ttest_stats[1]) +
  ylab("densité") +
  stat_function(fun = dnorm, col = "blue")

```

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 1.5. 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 pertinente (une différence de 0.28 euros sur un prix moyen de 82.56 euros est quantité négligeable).

1.2 Analyse exploratoire de données

Mieux vaut une réponse approximative à la bonne question, qu'une réponse exacte à la mauvaise question, qui peut toujours être précisée. — John Tukey

Avant d'ajuster un modèle aux données, il est souhaitable de comprendre leur nature pour éviter de mauvaises surprises et des erreurs d'interprétations. Cette section est inspirée du chapitre éponyme de l'ouvrage *R for Data Science* par Garrett Grolemund et Hadley Wickham. Une connaissance rudimentaire des graphiques est de mise et on s'attardera aux rudiments de la visualisation graphique. Je recommande également la lecture de la §~1.6 du livre *Introductory Statistics with*

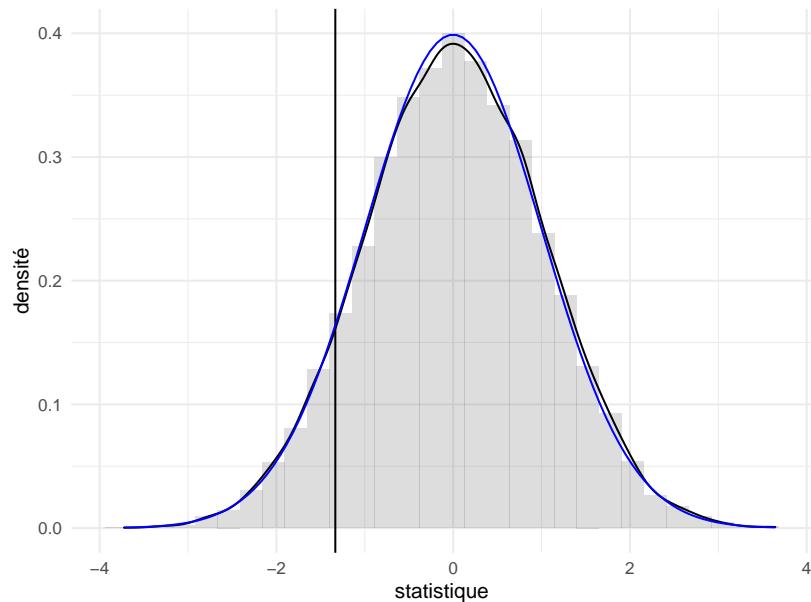


FIGURE 1.5 – Approximation par permutation de la loi nulle de la statistique de test de Welch (histogramme et trait noir) et loi asymptotique normale standard (trait bleu) pour le prix de billets de trains AVE au tarif promotionnel entre Madrid et Barcelone. La valeur de la statistique de test de l'échantillon original est représentée par un trait vertical.

Randomization and Simulation (en anglais) pour une revue. Si vous incluez un graphique (ou un tableau), il est important d'ajouter une légende qui décrit le graphique et le résume, de mettre les noms de variables (avec les unités) sur les axes, de soigner le rendu et le formatage pour obtenir un produit fini. Votre graphique raconte une histoire, aussi prenez-soin que cette dernière soit nécessaire et attrayante.

Si l'analyse exploratoire est souvent négligée dans les cours de statistique (parce qu'elle n'a pas de fondement mathématique), elle n'en est pas moins importante car elle nous sert à interpréter les données dans le contexte du problème et à nous assurer que notre analyse ou notre traitement de ces dernières est cohérent. Le sujet est difficile à cerner, puisque c'est davantage un art qu'une approche rigoureuse; Grolemund et Wickham parlent même « d'état d'esprit ». Le but de l'analyse exploratoire graphique est d'extraire des informations utiles, le plus souvent par le biais d'une série de questions qui sont raffinées au fur et à mesure que progresse l'analyse. On s'intéresse particulièrement aux relations et interactions entre différentes variables et la distribution empirique de chaque variable. Les étapes majeures sont :

1. Formuler des questions sur les données
2. Chercher des réponses à ces questions à l'aide de statistiques descriptives, de tableaux de

fréquence ou de contingence et de graphiques.

3. Raffiner nos questions, et utiliser les trouvailles pour peaufiner notre analyse

Dans un rapport, un résumé des caractéristiques les plus importantes devrait être inclus pour que le lecteur ou la lectrice puisse valider l'interprétation.

Commençons par les données : celles que l'on manipulera dans ce cours sont stockées sous forme tabulaire. Dans une base de donnée en format court, chaque ligne correspond à une observation et chaque colonne à une variable ; les entrées de la base de données contiennent les valeurs.

- 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.

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 ordinaires) si elles sont décrites par un adjectif; je préfère le terme catégorielle, plus évocateur.



FIGURE 1.6 – Illustration par Allison Horst de variables continues (gauche) et discrètes (droite).

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

- une variable discrète prend un nombre déterminé de valeurs ; ce sont souvent des variables de dénombrement ou des variables dychotomisées
- 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 (sexes, couleur, pays d'origine) ou ordinale (échelle de Likert, tranche salariale). La codification des modalités des variables catégorielles est arbitraire ; en revanche, on préservera l'ordre lorsqu'on représentera graphiquement les variables ordinaires. Lors de l'estimation, chaque variable catégorielle doit être transformée en un ensemble d'indicateurs binaires : 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.



FIGURE 1.7 – Illustration par Allison Horst de variables catégorielles nominales (gauche), ordinaires (centre) et binaires (droite).

Le principal type de graphique pour représenter la distribution d'une variable catégorielle est le diagramme à bande, 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 ordinaires. Cette représentation est en tout point supérieur au diagramme en camembert, une enseignance répandue 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 !

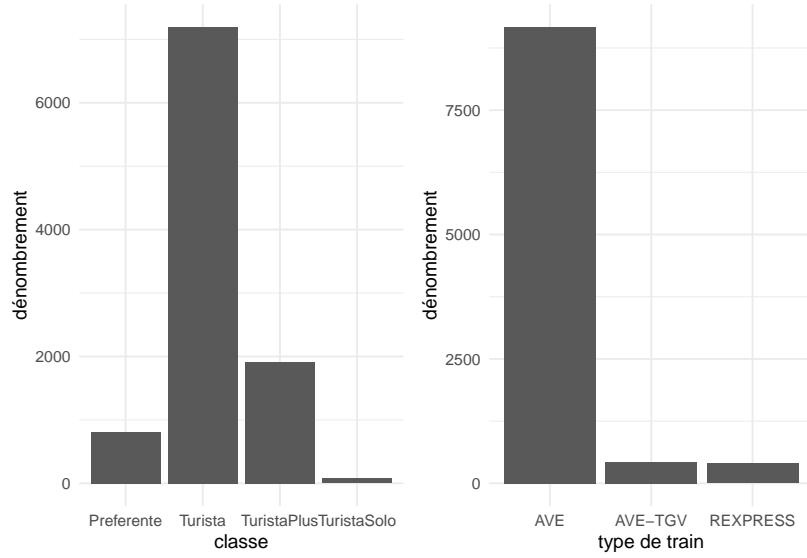


FIGURE 1.8 – Diagramme à bande pour la classe des billets de trains du jeu de données Renfe

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 discrétilisant 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 discrétilisée 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'asymétrie 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.

On peut voir en examinant l'étendue des données et en traçant un histogramme s'il y a des valeurs aberrantes ou inhabituelles. Une boîte-à-moustache (*boxplot*) représente graphiquement cinq statistiques descriptives.

- La boîte donne les 1e, 2e et 3e quartiles q_1, q_2, q_3 .
- 50% des observations sont au-dessus/en-dessous de la médiane q_2 .
- La longueur des moustaches est moins de 1.5 fois l'écart interquartile $q_3 - q_1$ (tracée entre

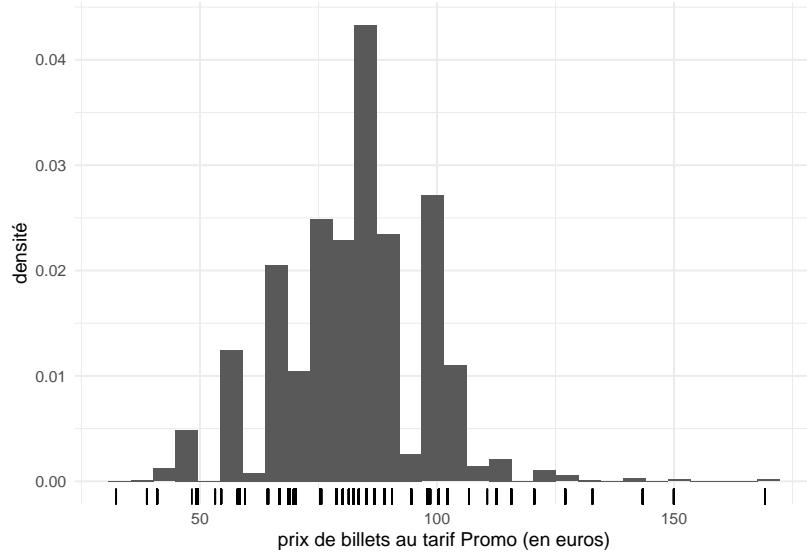


FIGURE 1.9 – Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe

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 augmentent. 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.

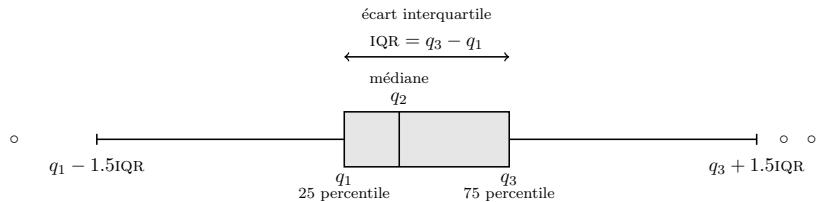


FIGURE 1.10 – Boîte à moustache

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 côté-à-côte. Une troisième variable catégorielle peut être ajoutée par le biais de couleurs, comme dans la Figure 1.11.

Si on veut représenter la covariate entre 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

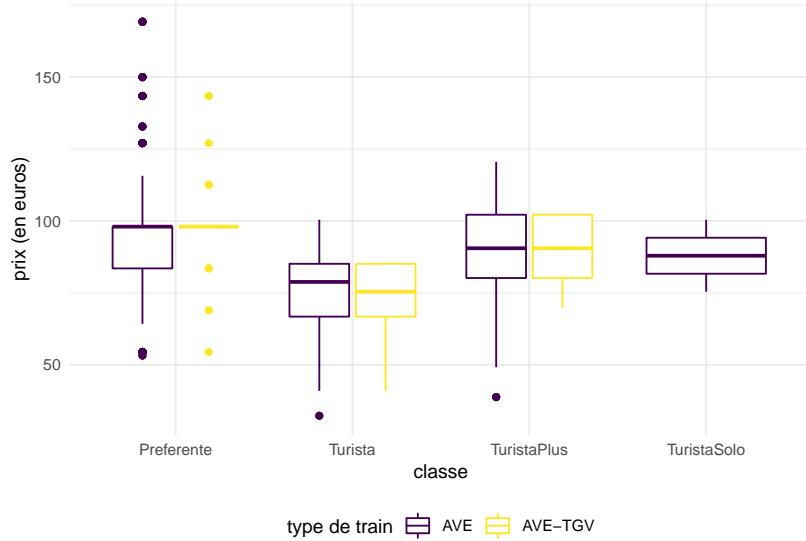


FIGURE 1.11 – Boîte à moustache du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe

des données peut être utile : vous verrez souvent des données positives à l'échelle logarithmique.

Par exemple, la Figure 1.12 montre le temps de réaction d'individus participant à une expérience effectuée au Tech3Lab ; on voit que le temps de réaction pour une personne qui texte est plus lent que lors d'une conversation téléphonique. La corrélation linéaire entre ces données appariées est faible.

Plutôt que de décrire plus en détail le processus de l'analyse exploratoire, on présente un exemple qui illustre le cheminement habituel sur les données de trains de la Renfe introduites précédemment.

Exemple 1.1 (Analyse exploratoire des trains Renfe). La première étape consisterait à lire la description de la base de données. Le jeu de données `renfe` contient les variables suivantes

- `prix` : prix du billet (en euros) ;
- `dest` : indicateur binaire du trajet, soit de Barcelone vers Madrid (0) ou de Madrid vers Barcelone (1) ;
- `tarif` : variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible` ;
- `classe` : classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo` ;
- `type` : variable catégorielle indiquant le type de train, soit Alta Velocidad Española (AVE), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et

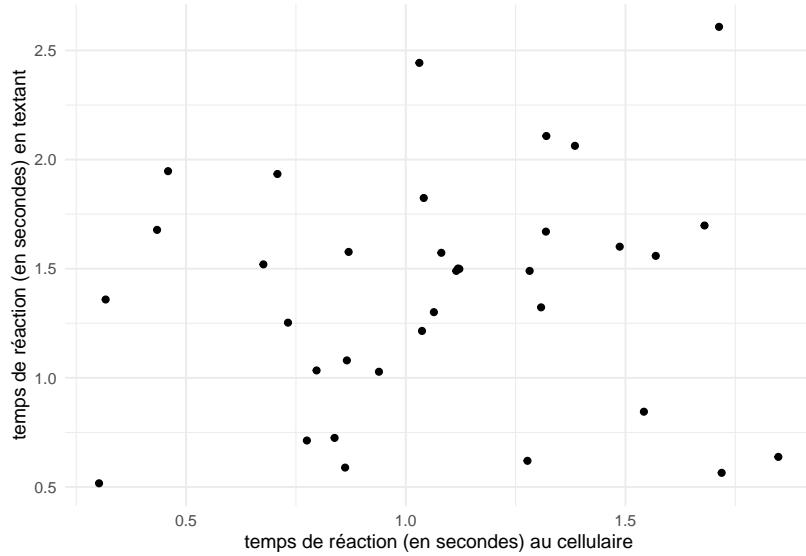


FIGURE 1.12 – Nuage de points du temps de réaction à deux tâches (données appariées) pour les données distractions

Renfe pour les trains à destination ou en provenance de Toulouse) AVE-TGV, soit les trains régionaux REXPRESS; seuls les trains étiquetés AVE ou AVE-TGV sont des trains à grande vitesse.

- duree : longueur annoncée du trajet (en minutes);
- jour entier indiquant le jour de la semaine du départ allant de dimanche (1) à samedi (7).

Il n'y a pas de valeurs manquantes et un aperçu des données (`head(renfe)`) montre qu'elles sont en format large, ce qui veut dire que chaque ligne correspond à un billet de train. On entame l'analyse exploratoire avec des questions plutôt vagues, par exemple

1. Quels sont les facteurs déterminant le prix et le temps de parcours ?
2. Est-ce que le temps de parcours est le même pour tous les types de train ?
3. Quelles sont les caractéristiques distinctives des types de train ?
4. Quelles sont les principales différences entre les tarifs ?

À l'exception de `prix` et de `duree`, toutes les variables explicatives sont catégorielles, parfois appelé facteur (`factor`). S'il faut déclarer chacune de ces variables, on porte une attention particulière à `jour` pour éviter les mauvaises surprises ultérieures.

On peut utiliser `str` pour obtenir un aperçu des données; la fonction `summary` permet d'obtenir des statistiques descriptives (minimum, maximum, moyenne, quartiles, nombre de valeurs manquantes) pour des variables continues et autrement la fréquence de variables catégorielles. On

peut créer une base de données pour chaque fréquence. La manipulation de variables dans des bases de données est parfois loin d'être toujours élégante en R : on accès aux variables avec \$, par exemple `renfe$prix`. Une alternative plus lisible et modulaire est d'utiliser l'opérateur tuyau (%>%), qui permet de créer une chaîne logique de commandes; la fonction `count` sert à compter le nombre d'instance de chaque modalité (ces fonctionnalités ne sont pas disponibles dans R par défaut, mais avec les paquetages tidyverse ou l'alternative minimale poorman, préalablement chargée).

```
renfe %>% count(classe)

##      classe     n
## 1 Preferente 809
## 2 Turista    7197
## 3 TuristaPlus 1916
## 4 TuristaSolo   78

# un raccourci pour la même syntaxe
renfe %>% group_by(type) %>% tally()

##      type     n
## 1 AVE 9174
## 2 AVE-TGV 429
## 3 REXPRESS 397

renfe %>% group_by(tarif) %>% tally()

##      tarif     n
## 1 AdultoIda 397
## 2 Flexible 1544
## 3 Promo 8059
```

En analysant le nombre de trains dans les catégories, on remarque qu'il y a autant de billets de type REXPRESS que le nombre de billets au tarif AdultoIda. On peut faire le décompte par catégorie avec un tableau de contingence, qui compte le nombre respectif dans chaque sous-catégorie. Dans la base de données Renfe, tous les billets pour les RegioExpress sont vendus au tarif AdultoIda en classe Turista. Le nombre de billets est minime, à peine 397 sur 10000. Cela suggère une nouvelle question : pourquoi ces trains sont-ils si peu populaires?

```
##      tarif     type     n
## 1 AdultoIda REXPRESS 397
```

```
## 2 Flexible      AVE 1446
## 3 Flexible     AVE-TGV   98
## 4    Promo       AVE 7728
## 5    Promo     AVE-TGV   331
```

On remarque également que le seul 17 temps de parcours sont affichés sur les billets (`renfe %>% distinct(duree)` ou `unique(renfe$duree)`). On peut donc penser que la durée affichée sur le billet (en minutes) est le temps de trajet annoncé. La majeure partie (15 sur 17) des temps de parcours sont sous la barre des 3h15, hormis deux qui dépassent les 9h! Selon Google Maps, les deux villes sont distantes de 615km par la route, 500km à vol d'oiseau. Cela implique que, vraisemblablement, certains trains dépassent les 200km/h, tandis que d'autres vont plutôt à 70km/h. Quels sont ces trains plus lents? La variable type codifie probablement ce fait, et permet de voir que ce sont les trains RegioExpress qui sont dans cette catégorie.

```
renfe %>%
  subset(duree > 200) %>%
  group_by(type, dest) %>%
  summarise("durée moyenne" = mean(duree),
            "écart-type" = sd(duree),
            "prix moyen" = mean(prix),
            "écart-type" = sd(prix))

##           type      dest durée moyenne écart-type prix moyen écart-type
## 1 REXPRESS Barcelone-Madrid      544          0     43.2        0
## 2 REXPRESS Madrid-Barcelone     562          0     43.2        0
```

Aller de Madrid à Barcelone à l'aide d'un train régulier prend 18 minutes de plus. Avec plus de 9h de trajet, pas étonnant donc que ces billets soient peu courus. Encore plus frappant, on note que le prix des billets est fixe : 43.25 euros peu importe que le trajet soit aller ou retour. C'est probablement la trouvaille la plus importante jusqu'à maintenant, car les billets de train de type RegioExpress ne forment pas un échantillon : il n'y a aucune variabilité! On aurait pu également découvrir cette anomalie en traçant une boîte à moustache du prix en fonction du type de train.

On pourrait soupçonner que les trains étiquetés AVE soient plus rapides, sachant que c'est l'acronyme de *Alta Velocidad Española*, littéralement haute vitesse espagnole. Qu'en est-il des distinctions entre les deux types de trains étiquetés AVE? Selon le site de la SNCF, les trains AVE-TGV sont des partenariats entre la Renfe et la SNCF et effectuent des liaisons entre la France et l'Espagne.

```
renfe %>%
  subset(type %in% c("AVE", "AVE-TGV")) %>%
  group_by(type, dest) %>%
```

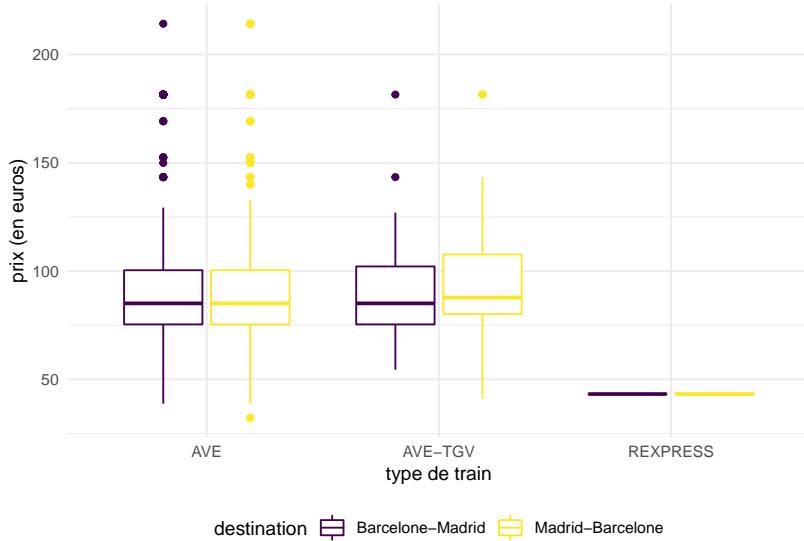


FIGURE 1.13 – Boîte à moustache du prix de billets de train de Renfe en fonction de la destination et du type de train.

```
summarise("durée moyenne" = mean(duree),
          "écart-type" = sd(duree),
          "prix moyen" = mean(prix),
          "écart-type" = sd(prix))
```

##	type	dest	durée moyenne	écart-type	prix moyen	écart-type
## 1	AVE	Barcelone-Madrid	171	15.9	87.4	19.8
## 2	AVE	Madrid-Barcelone	170	16.6	88.2	20.8
## 3	AVE-TGV	Barcelone-Madrid	175	0.0	87.0	16.8
## 4	AVE-TGV	Madrid-Barcelone	179	0.0	90.6	20.2

Les prix sont beaucoup plus élevés, en moyenne plus de deux fois plus que les trains régionaux. Les écarts de prix importants (l'écart type est de 20 euros) indique qu'il y a peut-être d'autres sources d'hétérogénéité, mais on pourrait soupçonner que la Renfe pratique la tarification dynamique. Il y un seul temps de parcours prévu pour les trains AVE-TGV. On ne note pas de différence de prix notable selon la direction ou le type de train grande vitesse, mais peut-être que les tarifs ou la classe disponibles diffèrent selon que le train ou non est en partenariat avec la compagnie française.

On a pas encore considéré le tarif et la classe des billets, hormis pour les trains RegioExpress. On voit dans la Figure 1.15 une forte différente dans l'hétérogénéité des prix selon le tarif; le tarif Promo prend plusieurs valeurs distinctes, tandis que les tarifs AdultoIda et Flexible semblent ne

prendre que quelques valeurs. La première classe (Preferente) est plus chère et il y a moins d'observations dans ce groupe. La classe Turista est la classe la moins dispendieuse et la plus populaire. TuristaPlus offre plus de confort, tandis que TuristaSolo permet d'obtenir un siège individuel.

Côté tarif, Promo et PromoPlus permettent d'obtenir des rabais pouvant aller jusqu'à respectivement 70% et 65%. Les annulations et changements ne sont pas possibles avec Promo, mais disponibles avec PromoPlus moyennant une pénalité équivalente à 30-20% du prix du billet. Le tarif Flexible est disponible au même prix que les billets réguliers, avec des bénéfices additionnels.

```
renfe %>% subset(tarif != "AdultoIda") %>%
  ggplot(aes(y = prix, x = classe, col = tarif)) +
  geom_boxplot() +
  labs(y = "prix (en euros)",
       x = "classe",
       color = "tarif") +
  theme(legend.position = "bottom")
```

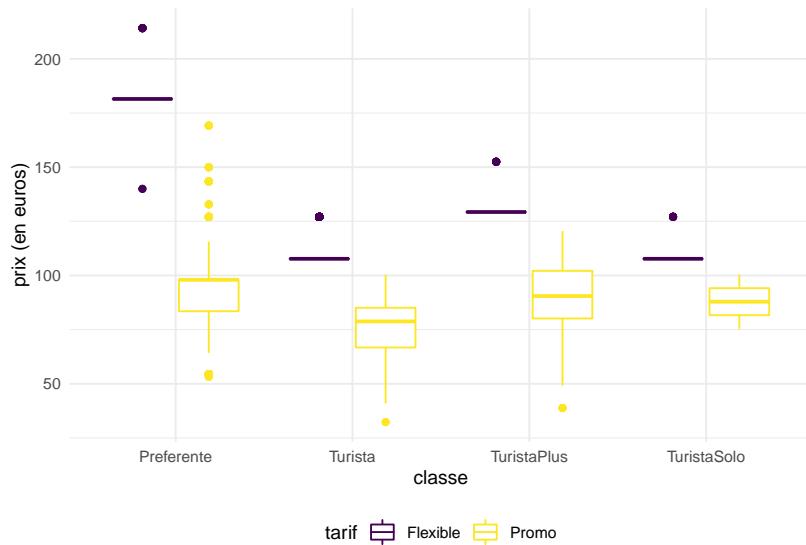


FIGURE 1.14 – Boîte à moustache du prix en fonction du tarif et de la classe de billets de trains à haute vitesse de la Renfe.

```
ggplot(data = renfe, aes(x = prix, y=..density.., fill = tarif)) +
  geom_histogram(binwidth = 5) +
  labs(x = "prix (en euros)", y = "densité") +
```

```
theme(legend.position = "bottom")
```

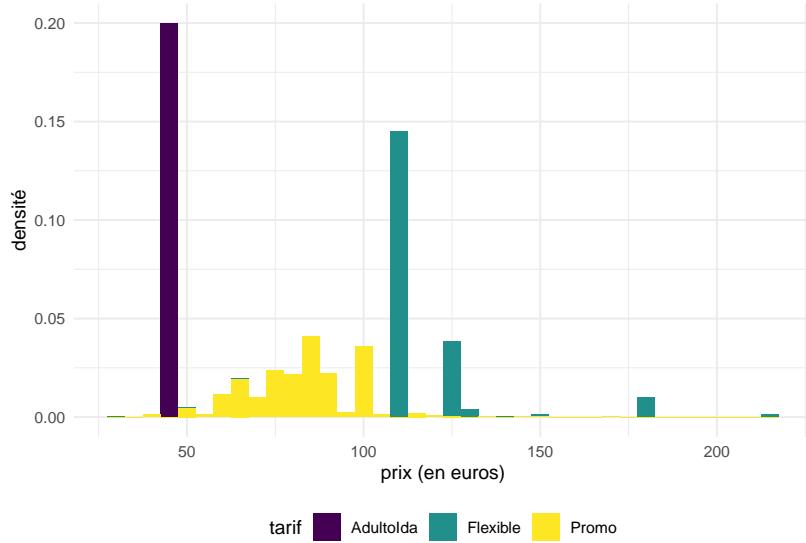


FIGURE 1.15 – Histogrammes du prix en fonction du tarif de billets de trains de la Renfe.

```
# Vérifier la répartition des billets Flexible
renfe %>% subset(tarif == "Flexible") %>% count(prix, classe)
```

```
##   prix      classe     n
## 1 108      Turista 1050
## 2 108  TuristaSolo    67
## 3 127      Turista  285
## 4 127  TuristaSolo     9
## 5 129  TuristaPlus    31
## 6 140 Preferente      2
## 7 152  TuristaPlus    10
## 8 182 Preferente     78
## 9 214 Preferente     12
```

On note que la répartition des prix pour les billets de classe Flexible est inhabituelle : notre boîte à moustache est écrasée et l'écart interquartile semble nul, même si quelques valeurs inexplicées sont aussi présentes. L'écrasante majorité des billets Flexibles sont en classe Turista, donc ça pourrait être dû à un (trop) faible nombre de billets dans chaque catégorie. On peut rejeter cette hypothèse en calculant le nombre de trains au tarif Flexible pour les différents types de billets. Ni la

durée, ni le type de train, ni la destination n'expliquent pas pourquoi le prix de certains billets Flexibles est plus faible ou élevés. Le prix des billets Promo est plus faible, et les billets au tarif Preferente (la première classe) sont plus élevés.

On peut résumer notre brève analyse exploratoire :

- plus de 91% des trains sont des trains à grande vitesse AVE.
- le temps de trajet dépend du type de train : les trains à grande vitesse mettent 3h20 au maximum pour relier Madrid et Barcelone.
- les temps de trajets sont ceux annoncés (variable discrète avec 17 valeurs uniques, dont 13 pour les trains AVE)
- le prix de trains RegioExpress est fixe (43.25€) ; tous ces billets sont dans la classe Turista et au tarif Adulto Ida. 57% de ces trains vont de Barcelone à Madrid. La durée du trajet pour les RegioExpress est de 9h22 de Barcelona à Madrid, 18 minutes de plus que dans l'autre direction.
- les billets en classe Preferente sont plus chers et moins fréquents. La classe Turista est la classe la moins dispendieuse et la plus populaire. Turista Plus offre plus de confort, tandis que Turista Solo permet d'obtenir un siège individuel.
- selon le site web de la Renfe, les billets au tarif Flexible « viennent avec des offres additionnelles qui permettent aux passagers d'échanger leurs billets ou annuler s'ils manquent leurs trains. » ; en contrepartie, ces billets sont plus chers et leur tarif est fixe sauf une poignée de billets dont le prix reste inexpliqué.
- la distribution des prix des billets de TGV au tarif Promo est plus ou moins symétrique, tandis que les billets au tarif Flexible apparaissent tronqués à gauche (le prix minimum pour ces billets est 107.7€ dans l'échantillon).
- la Renfe pratique la tarification dynamique pour les billets au tarif promotionnel Promo : ces derniers peuvent être jusqu'à 70% moins chers que les billets à prix régulier lorsqu'achetés via l'agence officielle ou le site de Renfe. Ces billets ne peuvent être ni remboursés, ni échangés.
- il n'y a pas d'indication à effet de quoi les prix varient selon la direction du trajet. “

Chapitre 2

Régression linéaire

Chapitre 3

Modèles linéaires généralisés

Chapitre 4

Données corrélées et longitudinales

Chapitre 5

Modèles linéaires mixtes

Chapitre 6

Analyse de survie

Chapitre 7

Inférence basée sur la vraisemblance

R

Je suis partisan de la philosophie Tinyverse; voir l'introduction sur les fonctionnalités de **R** avec des ressources et des dépendances minimales.

Annexe A

Compléments mathématiques

A.0.1 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. 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 tirés de la population à partir duquel on fait une analyse statistiques. La création de plans d'enquête 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. On doit ainsi éviter les biais de sélection, notamment les échantillons de commodité qui consistent en une sélection d'amis et de connaissances.

Notre échantillon est **aléatoire**, donc 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 notre échantillon sera grande, plus on obtiendra une mesure précise, voire exacte si on fait un recensement, de la quantité d'intérêt. L'exemple suivant illustre pourquoi le choix de l'échantillon est important.

Exemple A.1. 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. L'histoire complète (en anglais).

A.1 Variables aléatoires

Un phénomène d'intérêt varie d'un individu à l'autre (autrement un modèle statistique est rarement adéquat). On cherche à décrire son comportement, soit l'ensemble des valeurs possibles et leur probabilité/fréquence relative au sein de la population décrites par la loi de probabilité de la variable aléatoire.

On fera la distinction entre deux cas de figure : quand le phénomène prend des valeurs finies, comme par exemple un événement binaire (achat/non-achat d'un produit) ou un continuum de valeurs (par exemple, le prix d'un item). On dénote les variables aléatoires par des lettres majuscules : par exemple, $Y \sim \text{No}(\mu, \sigma^2)$ indique que Y suit une loi normale de paramètres μ et σ , qui représentent respectivement l'espérance et la variance de Y .

La fonction de répartition $F(y)$ donne la probabilité cumulative qu'un événement n'excède pas une variable donnée, $F(y) = \Pr(Y \leq y)$.

Si la variable Y prend des valeurs discrètes, alors on utilise la fonction de masse $f(y) = \Pr(Y = y)$ 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; la densité sert à estimer la probabilité que la variable Y appartienne à un ensemble B , via $\Pr(Y \in B) = \int_B f(y)dy$; la fonction de répartition est ainsi $F(y) = \int_{-\infty}^y f(x)dx$. Plusieurs lois aléatoires décrivent des phénomènes physiques simples et ont donc une justification empirique; on revisite les distributions les plus fréquemment couvertes.

Exemple A.2 (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 π . 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 $\Pr(Y = 1) = \pi$ et $\Pr(Y = 0) = 1 - \pi$ (complémentaire). La fonction de masse de la loi Bernoulli s'écrit de façon plus compacte

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

Un calcul rapide montre que $E(Y) = \pi$ et $Va(Y) = \pi(1 - \pi)$. 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 ?

Exemple A.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 $\text{Bin}(m, \pi)$; sa fonction de masse est

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

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

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 A.4 (Loi géométrique). La loi géométrique décrit le comportement du nombre d'essais Bernoulli de probabilité de succès π nécessaires avant l'obtention d'un premier succès. La fonction de masse de $Y \sim \text{Geo}(\pi)$ est

$$\Pr(Y = y) = \pi(1 - \pi)^{y-1}, \quad y = 1, 2, \dots$$

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 A.5 (Loi de Poisson). Si la probabilité d'un événement Bernoulli est **rare** dans le sens où $n\pi \rightarrow \lambda$ quand le nombre d'essais n augmente, alors le nombre de succès suit une loi de Poisson de fonction de masse

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

où $\Gamma(\cdot)$ dénote la fonction gamma. 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 $E(Y) = \text{Va}(Y) = \lambda$.

Exemple A.6 (Loi binomiale négative). On considère une série d'essais Bernoulli de probabilité de succès π 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

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

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 $Y | \Lambda = \lambda \sim \text{Po}(\lambda)$ 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 : on utilise la fonction de masse est

$$\Pr(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y, \quad y = 0, 1, \dots, \mu, r > 0,$$

où Γ dénote la fonction gamma. À noter que le paramètre $r > 0$ n'est plus nécessairement entier. La moyenne théorique et la variance sont $E(Y) = \mu$ et $\text{Va}(Y) = \mu + k\mu^2$, 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.

A.2 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. La loi des grands nombres établit que la moyenne empirique de n observations indépendantes de même espérance, \bar{Y}_n , tend vers l'espérance commune des variables μ , où $\bar{Y}_n \rightarrow \mu$. 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 utiles dans les expériences Monte Carlo : on peut ainsi approximer par simulation la moyenne d'une fonction de variables aléatoires $g(Y)$ 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.

A.3 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 \bar{Y} sera centré autour de μ ,
- l'erreur-type sera de σ/\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 \bar{Y} sera normale.

Mathématiquement, le théorème central limite dicte que $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \text{No}(0, \sigma^2)$. Si n est grand (typiquement supérieur à 30, mais cette règle dépend de la loi sous-jacente de Y), alors $\bar{Y} \sim \text{No}(\mu, \sigma^2/n)$.

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 (voir A.1) montre la distribution empirique des données.

Considérons maintenant des échantillons de taille dix. Dans notre premier échantillon aléatoire, la durée moyenne affichée est 167.7 minutes, elle est de 169.2 minutes dans le deuxième, de 172.7 dans le troisième, et ainsi de suite.

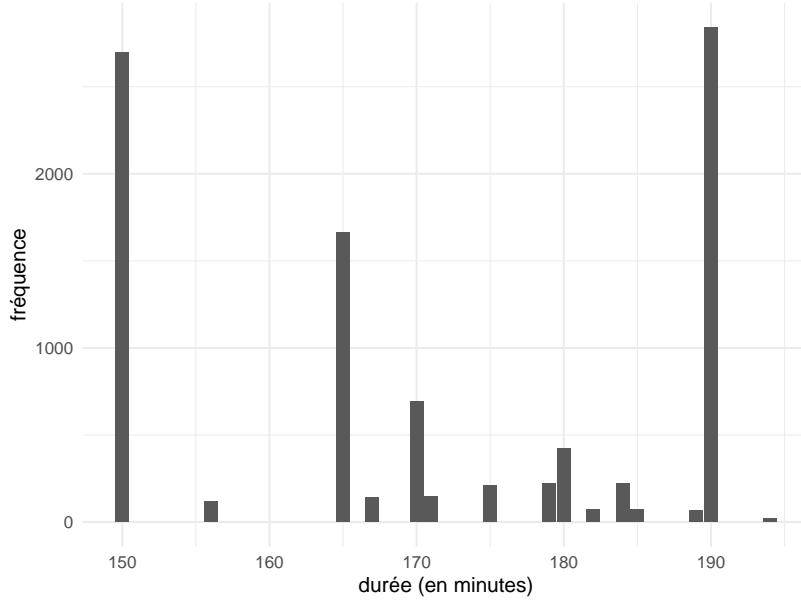


FIGURE A.1 – Distribution empirique des temps de trajet en trains à grande vitesse.

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 A.2 montre un de ces 1000 échantillons aléatoire de taille $n = 20$ tiré de notre base de données. Les autres graphiques de la Figure A.2 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 un seul échantillon dans l'exemple, mais vous pouvez vous amuser à regarder la taille de l'échantillon nécessaire pour que l'effet du théorème central limite prenne effet en simulant des observations d'une loi quelconque de variance finie. Cette applette permet de faire des simulations en variant la loi des données et la taille de l'échantillon.

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!

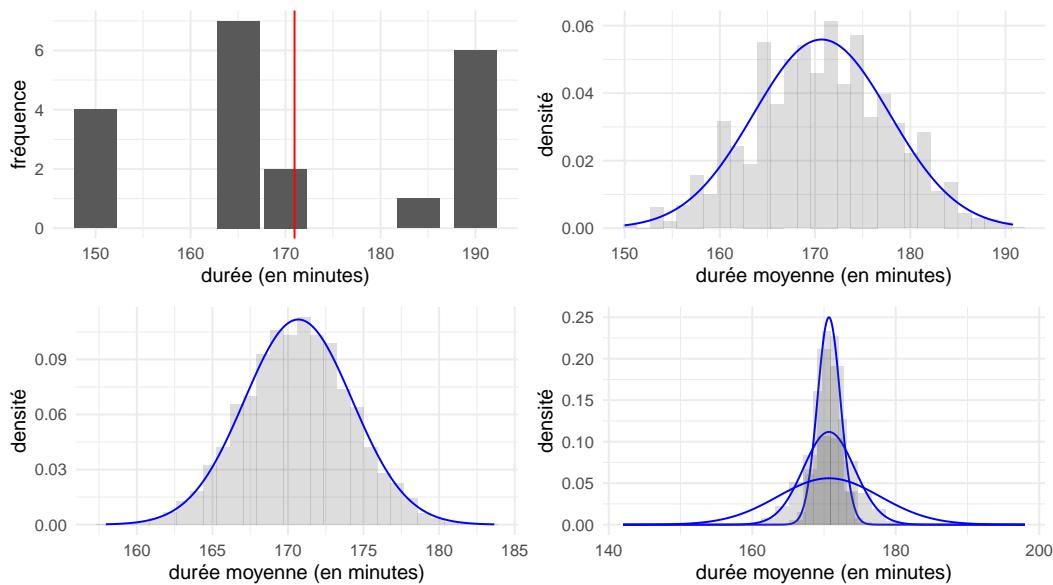


FIGURE A.2 – Représentation graphique du théorème central limite : échantillon aléatoire de 20 observations avec leur moyenne empirique (trait vertical rouge) (en haut à gauche). Les trois autres panneaux montrent les histogrammes des moyennes empiriques d'échantillons répétés de taille 5 (en haut à 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.

Bibliographie

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110–114.

Welch, B. L. (1947). The generalization of “Student’s” problem when several population variances are involved. *Biometrika*, 34(1–2), 28–35.