Statistiques avec R

SCI 1018

Tests d'hypothèse sur un seul groupe

Marc J. Mazerolle

Département des sciences du bois et de la forêt, Université Laval

Avec révisions mineures de

Élise Filotas, *Département science et technologie, Université TÉLUQ*, et

Marc-Olivier Martin-Guay, *Département des sciences biologiques, Université du Québec à Montréal*

Table des matières

1 Introduction

Lors de la leçon précédente, nous avons illustré le concept d'intervalle de confiance en nous appuyant sur le rééchantillonnage. Afin de construire des intervalles de confiance, nous avons utilisé la distribution normale ou la distribution du t de Student selon la taille de l'échantillon. Différentes stratégies d'échantillonnage ont été présentées, bien que l'échantillonnage complètement aléatoire et l'échantillonnage stratifié aléatoire soient les approches les plus fréquemment utilisées dans le cours.

2 Hypothèse

Une **hypothèse** est une explication potentielle que l'on formule pour décrire nos observations du monde extérieur. On entend par « observations », les « données » qui sont recueillies. Les hypothèses impliquent souvent une relation de cause à effet. Les études scientifiques permettent de comprendre la cause des phénomènes observés. La formulation d'hypothèses peut provenir d'observations personnelles préliminaires, de prédictions à partir de modèles théoriques, de la littérature scientifique ou du raisonnement.

Les données utilisées pour tester les hypothèses peuvent provenir d'**expériences contrôlées** (*controlled experiments*) ou d'**études d'observation** (*observational studies*). Lors d'une expérience contrôlée (exemple 3.1), les traitements sont assignés aléatoirement aux unités expérimentales, et on contrôle toutes les variables sauf celles pour lesquelles on veut mesurer l'effet. Le résultat d'une expérience est très robuste et permet de tirer des conclusions solides quant aux traitements étudiés. L'étude d'observation (exemple 3.2), quant à elle, implique de sélectionner aléatoirement les unités d'échantillonnage, mais l'assignation des traitements n'est pas déterminée aléatoirement par l'observateur. Puisqu'on contrôle très peu de variables, il est plus difficile de conclure en une relation de cause à effet d'une variable sur une autre.

On peut faire davantage confiance à des conclusions qui reposent sur des manipulations plutôt que sur des observations où on ne contrôle peu ou pas de variables. Les deux types

d'approches scientifiques sont importantes et complémentaires. L'expérience contrôlée peut parfois manquer de réalisme en simplifiant trop le système à l'étude, alors que l'étude d'observation mesure la variation en conditions plus naturelles, mais non contrôlées. Une étude d'observation peut révéler des patrons bruts et suggérer une expérience contrôlée afin de raffiner notre enquête.

Exemple 3.1 On réalise une expérience visant à caractériser l'effet d'une nouvelle crème médicale sur l'acné. Pour ce faire, on sélectionne 45 personnes volontaires, et on attribue aléatoirement à 15 d'entre elles la nouvelle crème, à 15 autres une crème sans la molécule active (placebo), et les 15 dernières ne reçoivent rien et servent de témoins. On s'assure que toutes les personnes participantes n'appliquent pas d'autres produits pouvant interférer avec les traitements et soumettent leurs peaux à des conditions similaires (p. ex., exposition au soleil modérée).

À la fin d'une période donnée pendant laquelle les personnes se sont appliquées une crème ou non, on évalue l'évolution de leurs problèmes de peau. L'expérience permettra ainsi de déterminer si l'application de la nouvelle crème a un effet bénéfique et si cet effet est supérieur à la crème sans molécule active.

Exemple 3.2 On réalise une étude d'observation afin d'étudier le nombre de grenouilles présentant des malformations dans des étangs en milieux agricoles. On sélectionne aléatoirement 20 étangs dans des milieux à faible intensité agricole et 20 étangs additionnels dans des milieux à forte intensité agricole. Dans chaque

étang, on effectue des inventaires afin d'estimer le nombre de grenouilles qui ont des malformations à l'aide de méthodes qui tiennent compte de la probabilité de détection. On mesure également d'autres variables, telles que la taille de l'étang, la distance de chaque étang par rapport au champ agricole le plus près, et la concentration d'une substance toxique dans l'eau de chaque étang.

Une analyse statistique permet de déterminer comment le nombre de malformations varie selon les autres variables mesurées. Si le seul effet que l'on observe est que le nombre de malformations est plus élevé dans les étangs près des champs agricoles, on ne peut pas conclure avec certitude que cette augmentation est due à la proximité des champs puisque nous n'avons pas contrôlé toutes les variables. En fait, il se peut que certaines substances toxiques différentes de celles qu'on a mesurées, provenant des champs agricoles situés à proximité, soient amenées par ruissellement dans les étangs. Ainsi, la cause des malformations n'est pas nécessairement la proximité des champs agricoles, mais plutôt les produits toxiques qui y sont appliqués. Après avoir détecté un effet de la proximité, il serait approprié d'effectuer une expérience contrôlée en faisant varier certaines variables pour quantifier directement leurs effets.

2.1 Hypothèse scientifique

Une **hypothèse scientifique** est une hypothèse qui peut être testée avec des observations ou des résultats expérimentaux et qui permettront de la modifier ou de la rejeter. Toute bonne hypothèse scientifique devrait générer de nouvelles prédictions[1]. À noter aussi qu'un modèle peut appuyer une hypothèse particulière et que ce modèle mène à des prédictions. Une

1. Par prédiction, on entend une généralisation d'un phénomène qui permet de prédire le comportement d'une variable d'intérêt pour une condition donnée. Par exemple, si on effectue une étude sur la pression artérielle et la masse corporelle, on pourrait par la suite prédire la pression artérielle pour une masse corporelle de 50 kg.

bonne hypothèse scientifique ne devrait fournir qu'une seule prédiction qui ne provient pas d'explications alternatives. Finalement, une bonne hypothèse doit avoir la possibilité d'être rejetée ou **réfutée** (ou falsifiée) en présence de données qui la contredisent.

Une hypothèse qui n'a aucune chance d'être rejetée, pour laquelle il est impossible de récolter des données, n'est pas une hypothèse scientifique – on doit distinguer croyance et hypothèse. Par exemple, on pourrait considérer les énoncés suivants qui ne peuvent pas être réfutés et qui ne sont donc pas de vraies hypothèses (pour le moment, en tout cas ...) :

> Il y a de la vie sur Pluton.
>
> Les populations de yétis et de sasquatchs sont repoussées dans des milieux encore plus reculés à la suite du développment urbain.
>
> Les chiens rêvent à leurs meilleurs moments de la journée pendant la nuit.

En contrepartie, les énoncés suivants sont des hypothèses qui peuvent être potentiellement réfutées avec des données :

> Les nombreux comportements à risques des hommes de 18 à 25 ans sur la route entraînent un taux plus important de mortalité lié aux accidents.
>
> Un groupe de grande taille confère à des poissons une meilleure chance de survie face à un prédateur.
>
> Les ordinateurs de la compagnie XX ont une durée de vie plus grande que les ordinateurs manufacturés par la compagnie YY.

Pour chacune des hypothèses ci-dessus, on pourrait développer un protocole afin de tester l'hypothèse à partir d'une expérience contrôlée ou d'une étude d'observation. Les données récoltées permettraient de rejeter ou non l'hypothèse. Il faut garder à l'esprit que la formulation d'une bonne hypothèse est le point de départ d'un bon projet scientifique.

2.2 Méthode scientifique

La **méthode scientifique** est une approche utilisée pour élaborer des hypothèses selon des observations et des prédictions. Différents types de raisonnements peuvent être utilisés selon la méthode scientifique, notamment la **déduction** (exemple 3.3) et l'**induction** (exemple 3.4). Le raisonnement déductif consiste à tirer des conclusions (inférer) à partir de données ou modèles partant du **général pour arriver au spécifique**. Par opposition, le raisonnement inductif consiste à tirer des conclusions à partir de données ou modèles partant du **spécifique au général**.

Exemple 3.3 Voici un exemple de déduction :

1. Tous les résidents du Québec ont une assurance médicament.
2. J'ai échantillonné un résident du Québec.
3. Ce résident a une assurance médicament.

Les éléments 1 et 2 sont des **prémisses** et le troisième est la conclusion. Ensemble, ces trois éléments forment un **syllogisme**.

Exemple 3.4 Voici un exemple d'induction :

1. Ces 25 résidents ont une assurance médicament.
2. Ces 25 résidents ont été échantillonnés au Québec.
3. Tous les résidents du Québec ont une assurance médicament.

Avec la déduction, si les prémisses sont vraies, la conclusion a de très bonnes chances d'être correcte, alors qu'avec l'induction, si les prémisses sont vraies, la conclusion est probablement vraie, mais elle peut aussi être fausse. On utilise l'un ou l'autre de ces raisonnements pour construire des hypothèses, mais on peut aussi les utiliser de façon concomitante. La statistique est une approche inductive, puisqu'on tire des conclusions en partant du spécifique (l'échantillon) pour l'appliquer au plus général (la population). La figure 1 montre l'utilisation de l'induction dans la démarche scientifique.

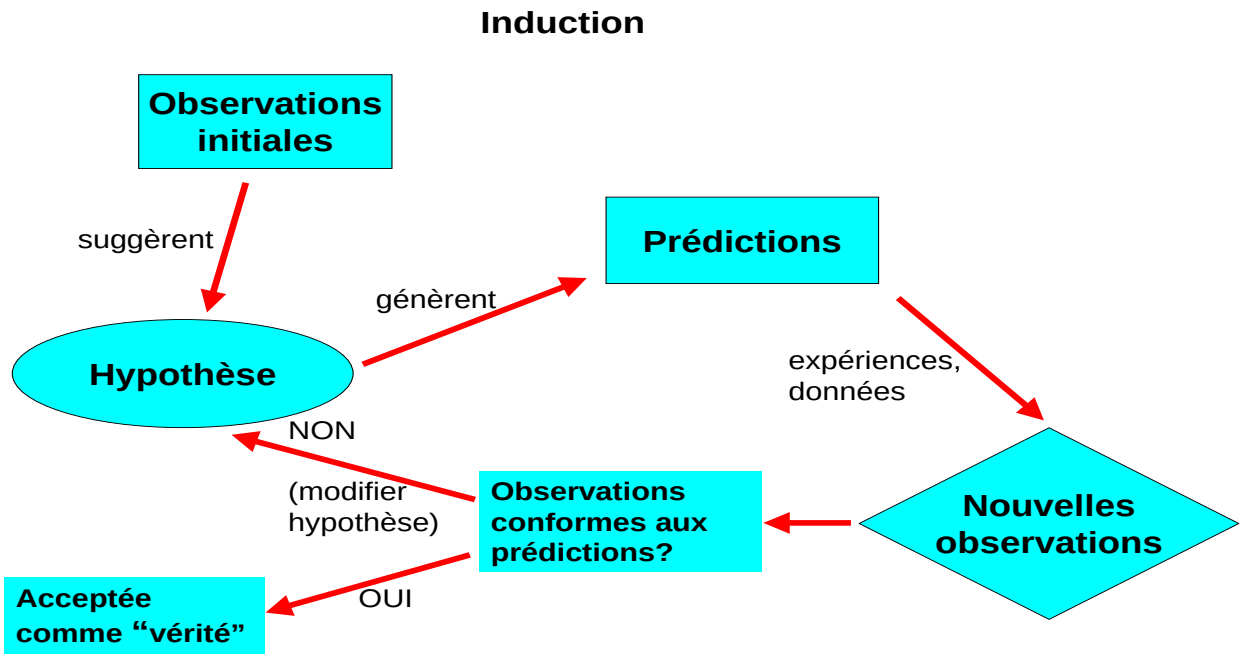

FIGURE 1 – Schéma de l'approche d'induction en sciences.

2.3 Inférence bayésienne

L'**inférence bayésienne** est une version moderne de l'approche inductive qui peut utiliser toute l'information disponible (résultats publiés, estimations des paramètres, croyances

ou connaissances) afin de construire une hypothèse. Cette méthode a été développée au 18e siècle. Cependant, comme les calculs sont très complexes, elle a été peu utilisée jusqu'au début des années 1990. Avec la venue d'ordinateurs puissants et d'algorithmes efficaces disponibles dans des logiciels libres (`BUGS`, `WinBUGS`, `OpenBUGS`, `JAGS`), cette branche des statistiques est en plein essor. Puisque le cours couvre les statistiques fréquentistes ou fishériennes, développées par Sir Ronald A. Fisher, nous ne discuterons pas plus des approches bayésiennes qui nécessitent une connaissance approfondie des statistiques.

2.4 Méthode hypothético-déductive

Avec la méthode hypothético-déductive, au lieu de commencer avec une seule hypothèse, on considère plusieurs hypothèses de travail pouvant expliquer le phénomène d'intérêt (fig. 2). Chaque hypothèse est testée et peut être réfutée avec de nouvelles données. On élimine des hypothèses au fur et à mesure selon les données qui sont récoltées dans des études ou des expériences successives. L'explication la plus plausible est l'hypothèse qui a résisté à la falsification à plusieurs reprises.

La méthode hypothético-déductive comporte plusieurs avantages :

- Elle force à considérer plusieurs hypothèses dès le début.
- Elle illustre les différences de prédiction entre chaque hypothèse.
- Les explications simples sont les premières considérées, les plus complexes ensuite.
- Plusieurs hypothèses sont testées en même temps (l'induction teste une hypothèse à la fois).

Toutefois, la méthode hypothético-déductive comporte un désavantage important :

- La méthode ne fonctionne pas si l'hypothèse correcte ne figure pas parmi les hypothèses énoncées avant l'expérience. En contrepartie, la méthode inductive peut commencer par une hypothèse incorrecte mais peut arriver à celle correcte après modification répétée de l'hypothèse suite à de nouvelles observations.

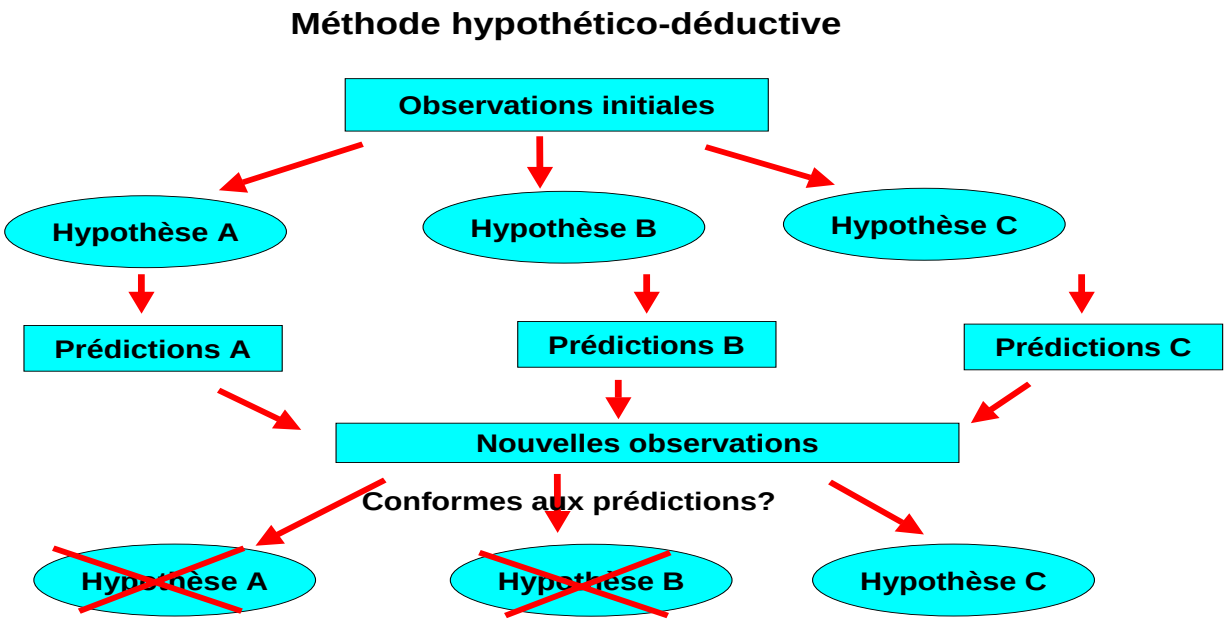

FIGURE 2 – Schéma de l'approche hypothético-déductive en sciences. L'hypothèse qui est conforme aux prédictions est celle qui a résisté à la falsification après plusieurs expériences successives.

3 Hypothèse statistique

Alors que l'hypothèse scientifique est l'élément qui devrait avoir motivé l'expérience, l'**hypothèse statistique** est une formulation « mathématique » de l'hypothèse scientifique. En effet, dans les approches classiques de tests d'hypothèse statistique, on compte deux hypothèses, l'**hypothèse nulle** (H_0) et l'**hypothèse alternative** (H_a). L'hypothèse nulle (H_0) représente l'absence de différence ou l'absence d'un effet. Elle se base sur un test statistique et une distribution indiquant que la variabilité observée dans les données est due au hasard ou à l'erreur d'échantillonnage. En d'autres mots, l'hypothèse nulle suppose qu'il ne se passe rien.

L'hypothèse alternative (H_a), quant à elle, est l'hypothèse statistique qui correspond à un effet ou une différence. C'est cette hypothèse qui indique qu'il se passe quelque chose. Le test d'hypothèse statistique consiste à essayer de rejeter H_0. Si l'hypothèse nulle (H_0) est

fausse, on la rejette et on se tourne vers l'hypothèse alternative (H_a). On parle toujours en termes de H_0 : on rejette H_0 ou non. On ne peut pas accepter H_0, seulement échouer de la rejeter faute de preuves. C'est ce système binaire d'hypothèses que les chercheurs utilisent depuis près d'un siècle. Toutefois, des méthodes alternatives ont été développées récemment qui permettent de comparer plusieurs hypothèses (> 2) simultanément, à l'aide de la sélection de modèles.

Exemple 3.5 On réalise une expérience pour tester l'hypothèse scientifique suivante : l'activité physique chez les élèves du secondaire influence les résultats scolaires.

Pour ce faire, on pourrait comparer les notes des élèves dans 10 classes qui ont une période supplémentaire d'éducation physique par semaine avec 10 autres classes sans la période supplémentaire (témoins). Après avoir récolté les données, on pourrait procéder à une analyse statistique pour tester les hypothèses statistiques suivantes à propos des moyennes des deux groupes :
H_0 (hypothèse nulle) : $\mu_{\textit{période suppl.}} = \mu_{\textit{témoins}}$
H_a (hypothèse alternative) : $\mu_{\textit{période suppl.}} \neq \mu_{\textit{témoins}}$

Si le test statistique indique que la différence observée entre les classes avec plus d'activité physique et les classes témoins est peu probable pour un échantillon de même taille tiré à partir de la même population lorsque H_0 est vraie, nous allons rejeter H_0. En d'autres mots, on se base sur une distribution statistique qui reflète H_0 pour situer la valeur observée de la statistique du test obtenue à partir de l'échantillon. Si cette valeur se trouve dans les extrémités de la distribution (c.-à-d., dans les queues), on va considérer cette valeur comme peu plausible lorsque H_0 est vraie. On conclura alors qu'il y a une différence du succès scolaire

entre les classes ayant eu une période d'éducation physique supplémentaire comparativement aux classes témoins, ce qui confirmera notre hypothèse scientifique de départ.

Pour déterminer si une différence est **statistiquement significative** (*statistically significant*) entre deux groupes ou deux échantillons, on doit procéder à un test statistique. Le test statistique est le processus par lequel on évalue la probabilité d'observer la valeur d'une statistique lorsque H_0 est vraie pour une taille d'échantillon donnée dans une population. Par le fait même, nous effectuons un test d'hypothèse.

Lorsqu'on compare les moyennes de deux groupes, il est peu probable d'observer exactement les mêmes valeurs dans les deux groupes. Il faut alors déterminer si la différence que nous avons observée est due à la variabilité naturelle dans la population ou à un effet réel (une vraie différence). Pour ce faire, il faut connaître la variabilité (la variance, s^2) dans chaque groupe. Il est normal que deux échantillons tirés de la même population soient différents en raison de la variabilité naturelle de la population et de l'erreur d'échantillonnage. Il faut un moyen de décider si la différence observée est due au hasard ou à un effet réel.

Le **seuil de signification** (α, *significance level*) est justement un critère qui est utilisé depuis longtemps pour distinguer entre le résultat du hasard et d'un effet réel. Par convention, on utilise $\alpha = 0.01, 0.05$ ou 0.10. Si la probabilité cumulative de la statistique du test utilisé est inférieure ou égale au seuil de signification ($P \leq \alpha$), on rejette H_0. Ainsi, avec $P \leq 0.05$, on devrait observer une différence comme celle de l'échantillon lorsque H_0 est vraie dans moins de 5 % des cas, si on répétait l'expérience. Autrement dit, si pour les caractéristiques des données le phénomène (H_0) est rare (c.-à-d., $P \leq \alpha$), on rejette H_0. Par conséquent, on ne rejette pas H_0 si le phénomène (H_0) est fréquent pour les caractéristiques des données de notre échantillon.

Exemple 3.6

On poursuit l'exemple précédent sur l'influence de l'activité physique sur le succès scolaire. Comme mentionné plus tôt, nous considérons les deux hypothèses suivantes et un seul de signification de 0.05 :

$H_0 : \mu_{période\ suppl.} = \mu_{témoins}$

$H_a : \mu_{période\ suppl.} \neq \mu_{témoins}$

$\alpha = 0.05$

On effectue un test statistique qui détermine la probabilité que l'échantillon provienne d'une population où H_0 est vraie. Si d'après le test, on a une probabilité cumulative inférieure ou égale à 5 % ($P \leq 0.05$) d'observer une différence comme celle dans notre échantillon lorsque H_0 est vraie, on rejette H_0. Autrement dit, si $P \leq 0.05$, il est peu probable d'observer une différence comme celle de notre échantillon lorsque H_0 est vraie si on répète l'étude un grand nombre de fois.

La déclaration que le résultat du test est statistiquement significatif apporte peu d'information. Il est préférable d'ajouter l'estimation du paramètre pour chaque groupe, comme la moyenne $\bar{x}$ ainsi qu'une mesure de précision. Toutefois, le choix d'un seuil de signification (0.05, 0.01, 0.10) n'est pas sans conséquence – il influence notre probabilité de commettre des erreurs dans nos conclusions.

4 Erreurs de type I et II

Le choix du seuil de signification influence la probabilité d'arriver à la bonne conclusion. Si on fixe α à 0.05, on va rejeter H_0 si la probabilité d'observer une différence comme celle de l'échantillon (lorsque H_0 est vraie) est inférieure à 0.05. Même avec la meilleure des intentions, on va donc rejeter H_0 dans des cas où H_0 est vraie avec une probabilité égale à α. Bien qu'il

Tableau 1 – Représentation des différents scénarios lors d'un test d'hypothèse nulle (H_0).

		CONCLUSION À PARTIR DU TEST	
		On ne rejette pas H_0	**On rejette H_0**
RÉALITÉ	H_0 **vraie**	décision correcte	erreur de type I (α) (faux positif)
	H_0 **fausse**	erreur de type II (β) (faux négatif)	décision correcte (puissance $= 1 - \beta$)

existe des valeurs dans les extrémités d'une distribution lorsque H_0 est vraie, notre choix de traiter ces valeurs comme des cas peu probables fait en sorte qu'on rejette parfois H_0 alors que nous n'aurions pas dû la rejeter. Avec $\alpha = 0.05$, on va rejeter H_0 à tort 5 % du temps (1 fois sur 20). On appelle ce type d'erreur, l'**erreur de type I** (erreur α, *type I error*). C'est équivalent de déclarer un accusé « coupable » alors qu'il est « non coupable » (tableau 1).

De la même façon, à d'autres occasions, nous ne rejetterons pas H_0 lorsqu'elle est fausse et qu'elle aurait dû être rejetée. Il s'agit de **l'erreur de type II** (erreur β, *type II error*) – on déclare l'accusé « non coupable » alors qu'il est « coupable ». La probabilité de commettre une erreur de type II peut se calculer à partir de formules ou de simulations, mais elle dépend de la taille de l'effet que l'on veut observer, de la taille de l'échantillon, de la variance des données, de la formulation de H_0, du design expérimental et du type d'échantillonnage. Les calculs de la puissance ne seront pas présentés dans le cours.

4.1 Puissance

La **puissance**, définie comme étant $1 - \beta$, est la probabilité de rejeter correctement H_0 lorsqu'elle est fausse. La puissance dépend de la taille de l'effet, de la taille de l'échantillon, de la variabilité des données, de l'hypothèse testée et du dispositif expérimental. Cette quantité peut être calculée à l'aide de données préliminaires avant d'effectuer l'échantillonnage afin de nous guider quant à la taille d'échantillon nécessaire étant donné l'effet que l'on veut mesurer

(p. ex., les différences entre moyennes de groupes). Le calcul de la puissance doit s'effectuer *a priori* avec des données préliminaires. Le calcul *a posteriori* de la puissance, à partir des données de l'expérience ou de l'étude d'observation, est à proscrire.

Dans toute analyse, nous voulons que la puissance soit élevée (le plus près possible de 1). Les mesures suivantes permettent d'augmenter la puissance d'une analyse :

- augmenter l'erreur de type I – si α augmente, β diminue, et par conséquent, $1 - \beta$ augmente ;
- augmenter la taille d'échantillon ;
- minimiser la variance des données.

4.2 Relation entre les erreurs de type I et II

On doit viser de faibles erreurs de type I et II. Toutefois, les deux types d'erreur sont liées entre elles : lorsqu'on réduit un type d'erreur, l'autre augmente automatiquement. Le choix de réduire un type d'erreur plutôt qu'un autre dépend du type d'étude réalisée et de la nature de sa discipline. Par exemple, lorsqu'on teste les effets secondaires d'un médicament en pharmacologie, il est à l'avantage d'une compagnie pharmaceutique de fixer le seuil α à une valeur plus faible afin de réduire la probabilité de déclarer un médicament nocif, ce qui augmente l'erreur de type II. En biologie de la conservation, on veut réduire les chances de ne pas déclarer un effet sur une espèce en voie d'extinction lorsqu'il y a réellement un effet – pour ce faire, on augmentera l'erreur de type I en augmentant le seuil α.

Certains auteurs argumentent que l'erreur de type I est la plus sérieuse. L'erreur de type I est liée à une déclaration (basée sur le rejet de H_0) qu'un mécanisme complexe se produit dans le système étudié alors que ce n'est pas le cas, puisqu'on a incorrectement rejeté H_0. D'autres chercheurs construiront leur expérience en se basant sur les résultats erronés. L'erreur de type II, quant à elle, représente notre incapacité à rejeter H_0 alors que nous aurions dû la rejeter. Cependant, quelqu'un avec un meilleur dispositif expérimental ou avec plus de données pourra la rejeter plus tard. Dans le domaine médical ou environnemental, l'erreur

de type II peut avoir de graves répercussions. Le meilleur moyen de réduire simultanément les deux types d'erreur est d'augmenter la taille de l'échantillon.

5 Utilisation de tests d'hypothèse

Pour utiliser correctement les tests d'hypothèse, nous avons besoin de données qui proviennent d'un échantillon aléatoire. Toutefois, il faut être conscient que les tests H_0 ne font pas l'unanimité puisque des problèmes potentiels sont associés à leur utilisation. Le choix du seuil α, c'est-à-dire la probabilité d'erreur de type I, est arbitraire. Il a des conséquences sur les erreurs de type I et II, sur la puissance, et donc sur nos conclusions. Le choix du seuil α devrait se baser sur les conséquences associées à commettre une erreur de type I ou II.

Par exemple, le coût d'une erreur de type I, comme pour le verdict de coupable pour un innocent, a de sérieuses conséquences pour un accusé de meurtre dans un pays avec peine de mort. Dans ce cas, il est souhaitable de réduire l'erreur de type I afin de diminuer le risque de rejeter H_0 alors qu'elle n'aurait pas dû être rejetée. Par contre, pour un accusé d'excès de vitesse, les conséquences d'une erreur de type I sont beaucoup moindres. En effet, le verdict de culpabilité requiert peu de preuves et entraîne une amende. Dans ce cas, on peut se permettre d'augmenter l'erreur de type I.

L'utilisation d'un seuil α rigide est déconseillée. Certains déclarent un effet à $P = 0.05$, mais pas à $P = 0.051$, même si le phénomène est toujours présent. L'utilisation de tests H_0 pour des résultats évidents n'amène pas beaucoup d'information (test pour déterminer si un échantillon avec $\bar{x} = 20$ et $SE = 1.2$ provient d'une population avec $\mu = 2000$). Il faut faire attention à l'interprétation incorrecte de P. P n'indique pas la probabilité que H_0 soit vraie. En fait, lors d'une expérience, H_0 est vraie ou fausse. La probabilité P indique la plausibilité des données de l'échantillon quand H_0 est vraie. Plus cette valeur est faible (c.-à-d., $P < \alpha$), moins il y a de preuve que notre échantillon provienne d'une population où l'hypothèse nulle est vraie.

Il est important de faire la distinction entre un effet statistiquement significatif et un effet biologiquement significatif. Avec un échantillon suffisamment grand, on peut rejeter n'importe quelle H_0. Il est donc important de distinguer lorsque l'effet, bien que statistiquement significatif, n'a pas d'importance pratique ou scientifique. C'est à l'expérimentateur de décider si une différence statistiquement significative a une importance scientifique. Par exemple, nous effectuons la comparaison entre un nouveau traitement de chimiothérapie et un traitement conventionnel sur la mortalité de cellules cancéreuses. À la fin de l'expérience, nous constatons que le nouveau traitement a tué 1% de cellules cancéreuses de plus que le traitement conventionnel ($\bar{x}_{\text{nouveau}} - \bar{x}_{\text{conventionnel}} = 1\%$). Est-ce que le résultat justifie d'adopter la nouvelle chimiothérapie alors que ses effets secondaires sont plus néfastes que le traitement conventionnel ?

Le non-rejet de H_0 n'équivaut pas à l'absence d'un effet. Si la puissance est trop faible, il sera impossible de rejeter H_0 lorsqu'elle est fausse. De plus, après un grand nombre de tests H_0, il est normal de rejeter H_0 même lorsqu'elle est vraie. Avec $\alpha = 0.05$, le rejet d'une seule H_0 sur 20 tests effectués pourrait être le fruit du pur hasard et une conséquence directe du choix de α.

5.1 Conseils sur la présentation des résultats de tests d'hypothèse

La simple déclaration que l'hypothèse nulle a été rejetée amène peu d'information. Il est préférable d'au moins présenter la valeur du test statistique, les degrés de liberté qui y sont associés ainsi que la valeur exacte du P. Ainsi, les lecteurs pourront juger par eux-mêmes de la pertinence de vos conclusions, qu'ils utilisent ou non le même seuil de signification que vous. Afin de bonifier l'utilisation des tests H_0, il est fortement conseillé de joindre à tout test d'hypothèse les estimations des paramètres (p. ex., moyennes de groupes) ainsi que des mesures de précision de ces estimations, telles que des erreurs-types ou des intervalles de confiance. Ces valeurs sont nécessaires dans les méta-analyses, lesquelles sont des analyses effectuées lors d'une revue de littérature sur un thème particulier à partir des résultats

d'articles publiés.

6 Estimation de paramètres

Lorsque nous effectuons une analyse statistique, nous désirons parfois obtenir plus d'information que de simplement savoir s'il y a un effet ou non. Par exemple, quel est le pourcentage de gens qui cliquent sur une publicité web dont on veut évaluer le succès ? Jusqu'à quelle température peut-on soumettre une pièce de moteur possédant un nouvel alliage ? La régression linéaire, où on estime une ordonnée à l'origine (axe des y) et une pente est un exemple de ce problème.

On peut utiliser les intervalles de confiance pour mesurer l'incertitude de l'estimation. On réalise rapidement que les tests d'hypothèse et les intervalles de confiance sont liés : les deux utilisent un seuil α. Ainsi, un intervalle de confiance qui inclut 0 correspond à $H_0 : \mathit{paramètre} = 0$. Nous reviendrons sur l'estimation de paramètres au cours des prochaines leçons.

7 Tests d'hypothèse sur la moyenne d'un seul groupe

Les tests d'hypothèse effectués sur la moyenne d'un groupe permettent de déterminer si un échantillon appartient à une population donnée. Typiquement, nous utilisons le test t pour réaliser cette analyse. En d'autres mots, on teste la probabilité d'observer une valeur de t qui a la même valeur ou qui est supérieure à celle calculée à partir de notre échantillon si on refaisait l'expérience avec un grand nombre d'échantillons tirés d'une population où l'hypothèse nulle est vraie. Ce concept de « rééchantillonnage » explique d'où vient le nom de statistiques « fréquentistes ». Autrement dit, on base nos conclusions sur la fréquence du phénomène d'intérêt que l'on aurait observée si on avait récolté un grand nombre d'échantillons de la même taille que notre échantillon original à partir de la population où l'hypothèse nulle est vraie. Bien sûr, on travaille avec un seul échantillon (les données que l'on a récoltées), mais on

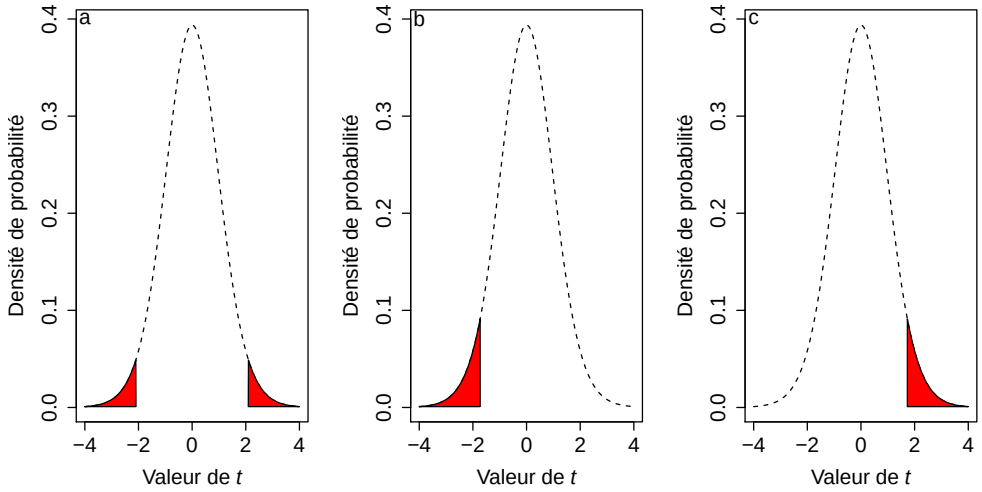

FIGURE 3 – Tests d'hypothèse bilatéral (a) et unilatéral (b, c) associés à la distribution du t de Student avec 19 degrés de liberté ($df = 19$). On note qu'avec un seuil $\alpha = 0.05$, on détermine si la valeur observée est dans le 2.5 % des queues de la distribution (0.025 à gauche + 0.025 à droite = 0.05). Le test unilatéral, quant à lui, détermine si la valeur observée se trouve dans le 5 % d'une des deux queues.

détermine la fréquence du phénomène qui nous intéresse si on avait eu plusieurs échantillons tirés d'une population en accord avec l'hypothèse nulle.

7.1 Test bilatéral et test unilatéral

Le test d'hypothèse peut être **bilatéral** (*two-sided test*) ou **unilatéral** (*one-sided test*). Dans le test bilatéral, on considère l'information contenue dans les deux queues de la distribution théorique (fig. 3a). À l'opposé, un test unilatéral s'intéresse spécifiquement à l'information contenue dans l'une des deux queues de la distribution (fig. 3b, c). Le choix d'exécuter un test unilatéral plutôt qu'un test bilatéral dépend des objectifs de l'analyse. Les deux prochains exemples illustrent le test unilatéral et le test bilatéral, respectivement.

Exemple 3.7

On effectue une étude en Montérégie afin de déterminer l'âge moyen d'une po-

pulation d'ours noirs que l'on veut comparer à la moyenne d'âge d'une autre population. On connaît l'âge moyen d'une population d'ours noirs en Abitibi, qui est de 7 ans.

On récolte un échantillon de $n = 20$ ours noirs en Montérégie et on détermine l'âge de chaque individu capturé. On peut ensuite effectuer un test d'hypothèse afin de savoir si l'âge moyen des ours qui ont été capturés en Montérégie diffère de l'âge moyen des ours noirs d'Abitibi. Ce test est bilatéral, puisqu'on s'intéresse à la fois à une différence positive ou négative entre les deux moyennes.

$H_0 : \mu = 7$ (hypothèse nulle ou de non-différence)

$H_a : \mu \neq 7$

$\alpha = 0.05$

Données récoltées (âges d'ours) :

18, 13, 17, 2, 8, 14, 8, 5, 9, 2, 8, 7, 7, 3, 19, 21, 20, 11, 7, 16

Si on teste sans spécifier si la moyenne des âges d'ours en Montérégie est supérieure ou inférieure à celle d'ours abitibiens, le test sera bilatéral. Nous pouvons utiliser le test t de Student pour nous assister dans notre décision de rejeter ou non H_0 :

$$t = \frac{\bar{x} - \mu}{SE} = \frac{10.8 - 7}{6.1/\sqrt{20}} = 2.748$$

On détermine ensuite la probabilité d'observer une valeur de t supérieure ou égale à celle qu'on a obtenue à partir de l'échantillon. Dans notre cas, on obtient $P(|t| \geq 2.748) = 0.0128$ – on a une probabilité de ~ 0.01 d'observer une valeur absolue de t supérieure ou égale à 2.748 dans une population où H_0 est vraie. En d'autres mots, il est peu probable d'observer cette valeur dans une population où H_0 est vraie.

Puisque nous avons fixé le seuil α à 0.05, et que la probabilité est inférieure à ce seuil, nous rejetons donc H_0. Étant donné le faible appui en faveur de cette hypothèse, nous concluons qu'il est peu probable que H_0 soit vraie. Nous pourrons donc dire que l'âge moyen des ours en Montérégie n'est pas de 7 ans.

Dans `R`, on peut trouver la solution en calculant le t à la main et en utilisant les fonctions telles que `qt( )` et `pt( )`. Ainsi, la solution dans `R` est :

```
> ##on crée un vecteur avec les observations
> age.ours <- c(18, 13, 17, 2, 8, 14, 8, 5, 9, 2, 8, 7, 7,
                3, 19, 21, 20, 11, 7, 16)
> ##on calcule la moyenne
> moy.age.ours <- mean(age.ours)
> ##on calcule l'erreur-type
> SE <- sd(age.ours)/sqrt(length(age.ours))
> ##on calcule le t
> t.val <- (moy.age.ours - 7)/SE
> t.val
[1] 2.747787
> ##on calcule les degrés de liberté
> df.ours <- length(age.ours) - 1
> ##on détermine P(t.val >= 2.748) - queue à droite
> p.droite <- 1 - pt(q = t.val, df = df.ours)
> ##on détermine P(t.val <= -2.748) - queue à gauche
> p.gauche <- pt(q = -t.val, df = df.ours)
> ##valeur du P dans les deux queues
> p.droite + p.gauche
```

```
[1] 0.0127962
> ##autre façon d'obtenir le P bilatéral
> ##puisque t est symétrique
> 2 * p.droite
[1] 0.0127962
```

Toutefois, on peut exécuter le test-t à l'aide de la fonction `t.test( )` et obtenir exactement le même résultat. Cette fonction comporte un argument `x` pour spécifier les valeurs de l'échantillon, un argument `mu` pour indiquer la valeur de la moyenne de la population à laquelle nous comparons la moyenne de l'échantillon (valeur de H_0), ainsi qu'un argument `alternative` qui donne l'hypothèse alternative testée, ici pouvant prendre les valeurs `"two.sided"` (bilatérale), `"less"` (inférieure à μ), `"greater"` (supérieure à μ).

```
> t.test(x = age.ours, mu = 7, alternative = "two.sided")

        One Sample t-test


data:  age.ours
t = 2.7478, df = 19, p-value = 0.0128
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
  7.893578 13.606422
sample estimates:
mean of x
    10.75
```

D'autres situations nous obligent à tester une hypothèse en dirigeant notre attention vers l'une des deux queues de la distribution, nous donnant ainsi un test unilatéral. C'est ce que nous montrons dans l'exemple suivant.

Exemple 3.8

On veut étudier l'efficacité d'un médicament sur la perte de poids. Pour ce faire, on sélectionne 12 patients chez qui on mesure la masse corporelle en kg. On administre un médicament qui stimule la perte de poids à chacun des 12 patients, à raison d'un comprimé par jour. Après un mois de la prise du médicament, on mesure le poids des mêmes individus. Les valeurs de l'échantillon montrent la différence entre le poids final et initial (kg). Une valeur positive indique un gain alors qu'une valeur négative est associée à une perte de poids : 0.2, -0.5, -1.3, -1.6, -0.7, 0.4, -0.1, 0.0, -0.6, -1.1, -1.2, -0.8

H_0 : $\mu \geq 0$ (il n'y a pas de perte de poids)

H_a : $\mu < 0$ (il y a perte de poids - à noter qu'un gain de poids indique que le médicament ne fonctionne pas)

$\alpha = 0.05$

On rejettera H_0 uniquement s'il est rare de rencontrer des valeurs aussi faibles ou plus faibles que celle de la statistique observée pour notre échantillon (dans la queue de gauche).

$$t = \frac{\bar{x} - \mu}{SE} = \frac{-0.608 - 0}{0.633/\sqrt{12}} = -3.329$$

```
> ##on crée un vecteur avec les observations
> poids <- c(0.2, -0.5, -1.3, -1.6, -0.7, 0.4,
             -0.1, 0.0, -0.6, -1.1, -1.2, -0.8)
> ##on calcule la moyenne
> moy.poids <- mean(poids)
> ##on calcule l'erreur-type
```

```
> SE <- sd(poids)/sqrt(length(poids))
> ##on calcule le t
> t.val <- (moy.poids - 0)/SE
> t.val
[1] -3.328513
> ##on calcule les degrés de liberté
> df.poids <- length(poids) - 1
> ##on détermine P(t.val <= -3.329) - queue à gauche
> p.gauche <- pt(q = t.val, df = df.poids)
```

On peut également utiliser `t.test( )` pour exécuter l'analyse de façon plus succincte :

```
> t.test(x = poids, mu = 0, alternative = "less")

        One Sample t-test


data:  poids
t = -3.3285, df = 11, p-value = 0.003364
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.2801098
sample estimates:
 mean of x
-0.6083333
```

On conclut que la probabilité d'observer une valeur de t de -3.329 dans un échantillon de 12 observations tiré d'une population où la moyenne est de 0 est très faible (c.-à-d., $P(t \leq -3.329) = 0.0034$). En d'autres mots, il est peu probable d'observer une valeur de t aussi faible que celle de l'échantillon lorsque la moyenne

de la population est de 0. Ainsi, nous rejetons H_0 et concluons que le médicament a stimulé la perte de poids.

7.2 Suppositions

Chaque analyse statistique comporte des **suppositions** (*assumption*) ou conditions qu'il faut respecter afin que les résultats du test soient valides. La vérification des suppositions fait partie de toute bonne analyse statistique. Si l'expérimentateur néglige cette étape importante, il sera impossible de faire la distinction entre des résultats pertinents et ceux qui devraient être mis à la poubelle. Le test-t sur un groupe suppose :

- que les données sont indépendantes, c'est-à-dire, qu'elles sont le résultat d'un échantillonnage aléatoire ;
- que les erreurs (résidus) suivent une distribution normale ;
- que la moyenne provient d'une distribution normale de moyennes.

Les deux premières suppositions **doivent** être vérifiées tandis qu'on ne peut pas vérifier formellement la troisième. Cependant, on peut considérer cette dernière comme valide, car elle découle du théorème de la limite centrale. Ce dernier indique que les moyennes d'échantillons indépendants tirés d'une même population constituent une distribution normale.

7.2.1 Indépendance des observations

L'**indépendance des observations** est une condition *sine qua non* à la réalisation des analyses classiques en statistiques. Si cette condition n'est pas respectée, la variabilité des données sera sous-estimée. De ce fait, on risque de rejeter plus facilement H_0 et d'arriver à des conclusions erronées. Toutes les analyses que nous voyons dans le cours, incluant celles des prochaines leçons, requièrent que l'indépendance des observations soit respectée.

À noter que des observations prises sur les mêmes unités, mais à des périodes différentes, ne sont pas indépendantes. Par exemple, mesurer la hauteur des mêmes arbres sur 5 an-

nées consécutives ou évaluer le succès scolaire des mêmes élèves sur plusieurs années ne sont pas des stratégies d'échantillonnage qui mènent à des observations indépendantes. Le meilleur moyen de s'assurer de l'indépendance des observations est d'utiliser une bonne stratégie d'échantillonnage. Certaines stratégies d'échantillonnage qui permettent d'obtenir un échantillon constitué d'observations indépendantes ont fait l'objet de la **leçon précédente**.

7.2.2 La normalité des résidus

On peut utiliser les **résidus** de l'analyse pour tester la normalité. On entend par résidus les erreurs ou la déviation des observations par rapport aux valeurs prédites par le modèle. Dans le cas du test-t, le modèle estime la moyenne de l'échantillon. Les résidus provenant du test-t sont donc la différence entre les valeurs observées (x_i) et les valeurs prédites ($\bar{x}$) :

$$\epsilon_i = x_i - \bar{x}$$

Dans l'exemple sur la perte de poids suite à la prise d'un médicament, on caculerait ainsi les résidus :

```
> ##on crée un vecteur avec les observations
> residus <- poids - mean(poids)
> residus
 [1]  0.808333333  0.108333333 -0.691666667 -0.991666667
 [5] -0.091666667  1.008333333  0.508333333  0.608333333
 [9]  0.008333333 -0.491666667 -0.591666667 -0.191666667
```

La vérification de la normalité peut se faire à l'aide de méthodes formelles telles que les tests d'Anderson-Darling, de Cramér-von Mises, ou de Shapiro-Wilk. L'hypothèse nulle de ces tests correspond au respect de la supposition de normalité (H_0 : la normalité est respectée). Le rejet de cette hypothèse indique que les résidus ne suivent pas la distribution normale. Plusieurs tests de normalité sont disponibles dans `R` dans le package (banque de fonctions) `nortest`.

Étant donné que ce package n'est pas distribué avec l'installation de base de R, celles et ceux ayant retenu l'approche par programmation doivent l'installer avant de pouvoir l'utiliser. Le document *Introduction à R - les packages* illustre toutes les étapes pour y arriver.

```
> ##on charge le package
> library(nortest)
> ##on fait le test d'Anderson-Darling
> ad.test(residus)
        Anderson-Darling normality test

data:  residus
A = 0.18668, p-value = 0.8815
> ##on fait le test de Cramer - von Mises
> cvm.test(residus)
        Cramer-von Mises normality test

data:  residus
W = 0.027575, p-value = 0.8641
> ##on fait le test de Shapiro-Wilk
> shapiro.test(residus)
        Shapiro-Wilk normality test

data:  residus
W = 0.96694, p-value = 0.8763
```

Dans ce cas, les trois tests suggèrent que les résidus suivent une distribution normale. Toutefois, chaque test a ses particularités. Certains sont sensibles à la trop faible ou à la trop grande taille de l'échantillon ou à la présence de valeurs extrêmes. Des méthodes graphiques informelles deviennent alors utiles, particulièrement pour des analyses statistiques

plus complexes comme nous verrons plus tard dans le cours. Par exemple, le graphique quantile-quantile est un outil efficace pour évaluer la normalité (fig. 4). On peut l'obtenir facilement dans `R` :

```
> ##un graphique quantile-quantile
> qqnorm(residus, main = "Graphique quantile-quantile",
        ylab = "Quantiles de l'échantillon",
        xlab = "Quantiles théoriques")
> ##on ajoute la droite théorique
> qqline(residus)
```

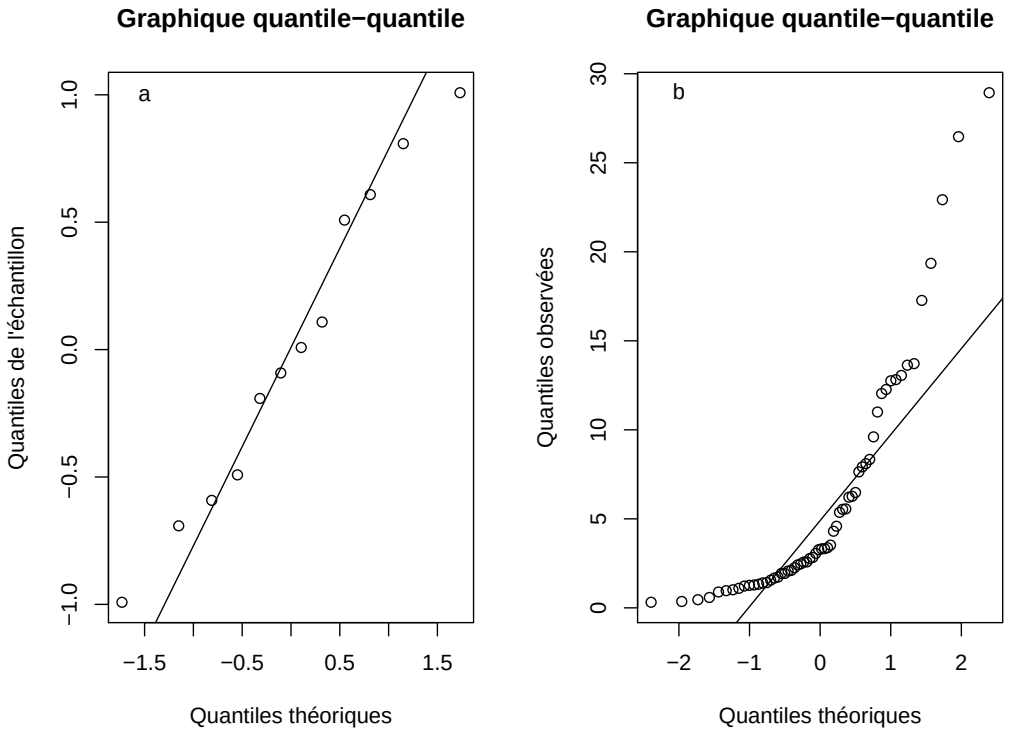

FIGURE 4 – Graphique quantile-quantile pour évaluer la normalité des résidus du test-t sur les données de perte de poids (a) et un échantillon qui ne suit pas la normalité (b).

Le graphique quantile-quantile compare les quantiles de la distribution des résidus à celle d'une distribution normale. On entend par quantile, par exemple, les percentiles des observations. Les quantiles observés sont comparés aux quantiles d'une distribution normale. Lorsque les quantiles observés correspondent à une distribution normale, on devrait voir une droite de 1 : 1 à 45°. L'ajout d'une droite théorique à l'aide de `qqline( )` permet d'évaluer si les quantiles observés ont une distribution normale. Si la plupart des points sont sur la droite, on peut conclure que les résidus ont une distribution normale. D'ailleurs c'est le cas des résidus de la perte de poids de notre exemple (fig. 4a). Le graphique de la figure 4b présente un autre échantillon et suggère que les valeurs ne suivent pas la distribution normale.

Conclusion

Dans ce texte, nous avons abordé la démarche scientifique, ainsi que l'élaboration d'hypothèses scientifiques et d'hypothèses statistiques. Nous avons ensuite présenté deux types d'erreur associées à l'exécution d'un test d'hypothèse : l'erreur de type I et l'erreur de type II. Le seuil de signification (α) détermine la probabilité de rejeter faussement une hypothèse nulle qui est vraie (erreur de type I), alors que la probabilité de ne pas rejeter une hypothèse nulle qui est fausse est donnée par β (erreur de type II). La puissance est la probabilité de correctement rejeter une hypothèse nulle qui est fausse et est donnée par $1 - \beta$. Le meilleur moyen d'augmenter la puissance d'un test est d'augmenter la taille d'échantillon et de minimiser la variance. Comme première application d'un test d'hypothèse statistique, nous avons présenté le test-t unilatéral et bilatéral sur un groupe, les suppositions sous-jacentes à ce test, ainsi que des outils permettant de vérifier ces suppositions.

Index