

TP de Biométrie Semestre 6

Benoît Simon-Bouhet

dimanche 27 août 2023

Table des matières

Introduction	4
Objectifs	4
Pré-requis	5
Organisation	6
Volume de travail	6
Modalités d'enseignement	7
Utilisation de Slack	9
Progression conseillée	11
Évaluation(s)	13
Licence	15
1 Comparaison de moyennes : plus de 2 groupes	16
1.1 Pré-requis	16
1.2 Contexte	17
1.3 Importation et mise en forme des données . .	18
1.4 Exploration statistique des données	21
1.5 Exploration graphique	23
1.6 Le test paramétrique	27
1.6.1 Réalisation du test	27
1.6.2 Conditions d'application	29
1.6.3 Interprétation des résultats	34
1.6.4 Tests <i>a posteriori</i> ou tests <i>post-hoc</i> . .	35
1.7 L'alternative non paramétrique	41
1.7.1 La robustesse de l'ANOVA	41
1.7.2 Réalisation du tests et interprétation .	42
1.7.3 Tests <i>a posteriori</i> ou tests <i>post-hoc</i> . .	43
1.8 Exercices d'application	47
1.8.1 <i>Cardamine pensylvanica</i>	47
1.8.2 Insecticides	48
1.8.3 La longueur des nageoires des man- chots femelles	48
2 Corrélation	49
2.1 Pré-requis	49
2.2 Principe	50
2.3 Contexte	53

2.4	Importation et mise en forme des données . . .	53
2.5	Exploration statistique des données	56
2.6	Exploration graphique des données	58
2.7	Le test paramétrique	60
2.7.1	Les hypothèses	60
2.7.2	Conditions d'application	61
2.7.3	Réalisation du test et interprétation .	64
2.7.4	Estimation et intervalle de confiance .	66
2.8	Corrélation et causalité	67
2.8.1	Quelques exemples évidents	67
2.8.2	Les variables confondantes	68
2.8.3	Études expérimentales ou observationnelles	69
2.9	L'alternative non paramétrique	72
2.10	Exercices	73
2.10.1	<i>Canis lupus</i>	73
2.10.2	Les miracles de la mémoire	74
3	Régression linéaire	76
3.1	Pré-requis	76
3.2	Principe	77
3.3	Contexte	79
3.4	Importation et mise en forme des données . .	79
3.5	Exploration statistique des données	80
3.6	Exploration graphique des données	82
3.7	Le test paramétrique	85
3.7.1	Les hypothèses	85
3.7.2	Réalisation du test	86
3.7.3	Conditions d'application	87
3.7.4	Interprétation des résultats	91
3.7.5	Intervalle de confiance de la régression	93
3.8	L'alternative non paramétrique	96
3.9	Exercices	97
3.9.1	Datasaurus et Anscombe	97
3.9.2	In your face	99
4	Comparaison de proportions	100
	References	101

Introduction

Objectifs

Ce livre contient l'ensemble du matériel (contenus, exemples, exercices...) nécessaire à la réalisation des travaux pratiques de **Biométrie** de l'EC '*Outils pour l'étude et la compréhension du vivant 5*' du semestre 6 de la licence Sciences de la Vie de La Rochelle Université.

À la fin du semestre, vous devriez être capables de faire les choses suivantes dans le logiciel **RStudio** :


- Explorer des jeux de données en produisant des résumés statistiques de variables de différentes nature (numériques continues ou catégorielles) et en produisant des graphiques appropriés
- Calculer des statistiques descriptives (moyennes, médianes, quartiles, écart-types, variances, erreurs standard, intervalles de confiance, etc.) pour plusieurs sous-groupes de vos jeux de données, et les représenter sur des graphiques adaptés
- Choisir et formuler des hypothèses adaptées à la question scientifique posée (hypothèses bilatérales ou unilatérales)
- Choisir les tests statistiques permettant de répondre à une question scientifique précise selon la nature de la question posée et la nature des variables à disposition
- Réaliser les tests usuels de comparaison de proportions et de moyennes (χ^2 , t de Student à 1 ou 2 échantillons, appariés ou indépendants, ANOVAs etc.) et d'association entre 2 variables numériques (régression linéaire)
- Vérifier les conditions d'application des tests, et le cas échéant, réaliser des tests non paramétriques équivalents
- Interpréter correctement les résultats des tests pour répondre aux questions scientifiques posées

Pré-requis

Pour atteindre les objectifs fixés ici, et compte tenu du volume horaire restreint qui est consacré aux TP et TEA de Biométrie au S6, vous devez impérativement posséder un certain nombre de pré-requis. En particulier, vous devriez avoir à ce stade une bonne connaissance de l'interface des logiciels R et RStudio, et vous devriez être capables :

1. de créer un **Rproject** et un script d'analyse dans RStudio
2. d'importer des jeux de données issus de tableurs dans RStudio
3. d'effectuer des manipulations de données simples (sélectionner des variables, trier des colonnes, filtrer des lignes, créer de nouvelles variables, etc.)
4. de produire des graphiques de qualité, adaptés à la fois aux variables dont vous disposez et aux questions auxquelles vous souhaitez répondre.
5. de calculer des indices de statistiques descriptives de position (moyenne, médiane, quartiles), de dispersion (variance, écart-type, intervalle interquartile) et de d'incertitude (erreur standard et intervalle de confiance), pour un échantillon ou pour chaque modalité d'un facteur
6. de représenter sur un graphique des données accompagnées des barres d'erreur pertinentes

Vous devriez en outre être capable d'expliquer les différences entre dispersion et incertitude, et de choisir le type de graphique adapté selon le nombre et la nature des variables dont vous disposez, et selon la question à laquelle vous tentez de répondre.

 Si ces pré-requis ne sont pas maîtrisés

Mettez-vous à niveau de toute urgence en lisant attentivement :

1. [le livre en ligne de Biométrie du semestre 3](#). Vous y trouverez les éléments de prise en main du logiciel, les explications concernant les représentations graphiques et la manipulation de tableaux de données dans RStudio.

2. [le livre en ligne de Biométrie du semestre 4](#). Vous y trouverez notamment les explications concernant les statistiques descriptives, les notions de position, de dispersion et d'incertitude.

Sans une bonne maîtrise de ces outils et notions, vous aurez du mal à suivre ce que nous allons aborder ce semestre.

Organisation

Volume de travail

Au total, 4 séances de TP d'1h30 suivies de 4 séances de TEA d'1h30 sont prévues entre le 27 février et le 24 mars 2023 :

- Semaine 09 (du 27 février au 3 mars octobre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 10 (du 06 au 10 mars) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 11 (du 13 au 17 mars) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 12 (du 20 au 24 mars) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30

Toutes les séances de TP ont lieu les lundis matins de 8h00 à 13h00 (3 groupes) et les vendredi après-midi de 13h15 à 16h30 (2 groupes). Il y a cependant 2 exceptions à cela : le vendredi 3 mars, les TP auront lieu de 15h00 à 18h15, et la semaine 12, les TP auront lieu le mardi matin de 8h00 à 13h00 au lieu du lundi matin.

Tous les TP ont lieu en salle MSI 217. Tous les TEA sont à distance.

Je vous indique tout ceci pour vous permettre de vous déplacer aux séances qui vous conviennent le mieux. Si vous êtes disponible pendant le TP d'un autre groupe mais que vous avez des questions à poser, n'hésitez pas à venir en MSI 217 : j'y serai systématiquement.

Au total, chaque groupe aura donc 4 séances de TP et 4 séances de TEA, soit un total de 12 heures prévues dans

vos emplois du temps. C'est peu pour atteindre les objectifs fixés et il y aura donc évidemment du travail personnel à fournir en dehors de ces séances. J'estime que vous devrez fournir à peu près une douzaine d'heures de travail personnel en plus des séances prévues dans votre emploi du temps. Sachez toutefois que selon votre niveau d'aisance, et les acquis des semestres précédents, vous pourrez aller plus vite que prévu, ou au contraire (beaucoup) plus lentement ! Attention donc : pensez bien à prévoir du temps dans vos plannings car le travail personnel est essentiel pour progresser dans cette matière. J'insiste sur l'importance de faire l'effort dès maintenant : vous aurez très probablement besoin de savoir faire ce qui est au programme de ce semestre pendant votre stage et, très vraisemblablement, dans vos futurs masters également. C'est donc maintenant qu'il faut acquérir des automatismes, cela vous fera gagner énormément de temps ensuite.

Modalités d'enseignement

Pour suivre cet enseignement vous pourrez utiliser les ordinateurs de l'université, mais je ne peux que vous encourager à utiliser vos propres ordinateurs, sous Windows, Linux ou MacOS. Lors de vos futurs stages et pour rédiger vos comptes-rendus de TP, vous utiliserez le plus souvent vos propres ordinateurs, autant prendre dès maintenant de bonnes habitudes en installant les logiciels dont vous aurez besoin tout au long de votre licence. Si vous n'avez pas suivi les enseignements de biométrie du semestre 5 et que les logiciels R et RStudio ne sont pas encore installés sur vos ordinateurs, suivez [la procédure décrite ici](#). Si vous ne possédez pas d'ordinateur, manifestez vous rapidement auprès de moi car des solutions existent (prêt par l'université, travail sur tablette via [RStudio cloud](#)...).

! Important

L'essentiel du contenu de cet enseignement peut être abordé en autonomie, à distance, grâce à ce livre en ligne, aux ressources mises à disposition sur Moodle et à votre ordinateur personnel. Cela signifie que **la présence physique lors de ces séances de TP n'est pas obligatoire**.

Plus que des séances de TP classiques, considérez plutôt qu'il s'agit de **permanences non-obligatoires** : si vous pensez avoir besoin d'aide, si vous avez des points de blocage ou des questions sur le contenu de ce document ou sur les exercices demandés, alors venez poser vos questions lors des séances de TP (les vôtres ou celles de vos collègues). Vous ne serez d'ailleurs pas tenus de rester pendant 1h30 : si vous obtenez une réponse en 10 minutes et que vous préférez travailler ailleurs, vous serez libres de repartir !

De même, si vous n'avez pas de difficulté de compréhension et les exercices de ce livre en ligne ne vous posent pas de problème, votre présence n'est pas requise. Si vous souhaitez malgré tout venir en salle de TP, pas de problème, vous y serez toujours les bienvenus. Mais attention, on a parfois l'impression d'avoir bien compris lorsque l'on lit des explications. La seule façon d'en être sûr, c'est d'être capable d'expliquer (un concept, une notion, un raisonnement, une méthode...) à une autre personne. Si c'est clair pour vous, vous devriez être en mesure d'expliquer facilement à une tierce personne, en utilisant le vocabulaire approprié. Donc si vous avez le moindre doute **faites l'effort de passer en salle de TP**, ne serait-ce que quelques minutes, afin de confirmer auprès de moi que vous avez bien compris. Cela ne vous coûte pas grand chose : les créneaux de TP apparaissent de toutes façons dans vos emplois du temps.

Ce fonctionnement très souple a de nombreux avantages :

- vous vous organisez comme vous le souhaitez
- vous ne venez que lorsque vous en avez vraiment besoin
- celles et ceux qui se déplacent reçoivent une aide personnalisée
- vous travaillez sur vos ordinateurs
- les effectifs étant réduits, c'est aussi plus confortable pour moi !

Toutefois, pour que cette organisation fonctionne, cela demande de la rigueur de votre part, en particulier sur la régularité du travail que vous devez fournir. Si la présence en salle de TP n'est pas requise, **le travail demandé est bel et bien obligatoire** ! Si vous venez en salle de TP sans avoir travaillé en amont, vous risquez de perdre votre temps car vous passerez votre séance à lire et suivre ce livre en ligne, choses que vous pouvez très bien faire chez vous. De même,

si vous attendez le 20 mars pour vous y mettre sérieusement, je ne pourrai pas grand chose pour vous. Je le répète, outre les heures de TP/TEA prévus dans vos emplois du temps, vous devez prévoir au moins 12 heures de travail personnel supplémentaire.

Pour finir sur cette question de l'organisation de ces TP et TEA, je vous livre un commentaire qui m'a été fait lors des semestres précédents lorsque je demandais à vos collègues et prédécesseurs s'ils appréciaient ou non ce fonctionnement. Plusieurs m'ont dit ceci : “nous aurions préféré des séances de présentiel classique car comme ça, on aurait pu vous poser nos questions directement”. J'avoue ne pas avoir su quoi répondre... Encore une fois, les créneaux sont prévus dans vos emplois du temps, je suis physiquement présent en salle de TP pour toutes les séances de tous les groupes. Donc si vous avez des questions à poser, et si c'est plus facile pour vous, venez comme à une séance de TP classique, rien ne vous en empêche !

Utilisation de Slack

Comme au semestre précédent, nous pourrons échanger sur [l'application Slack](#). Si vous ne l'avez pas encore fait, créez-vous un compte en ligne et installez le logiciel sur votre ordinateur (il existe aussi des versions pour tablettes et smartphones). Lorsque vous aurez installé le logiciel, [cliquez sur ce lien](#) pour vous connecter à notre espace de travail commun intitulé L3 SV 22-23 / EC outils (ce lien expire régulièrement : faites moi signe s'il n'est plus valide). C'est le même espace de travail qu'au semestre précédent et si vous vous y êtes déjà connecté cet automne, vous n'avez plus qu'à relancer le logiciel.

Vous verrez que 3 “canaux” y sont disponibles :

- #général : c'est là que les questions liées à l'organisation générale du cours, des TP et TEA, des évaluations, etc. doivent être posées. Si vous ne savez pas si une séance de permanence a lieu, posez la question ici.
- #questions-rstudio : c'est ici que toutes les questions pratiques liées à l'utilisation de R et RStudio devront être posées. Problèmes de syntaxe, problèmes liés à l'interface, à l'installation des packages ou à l'utilisation

des fonctions, à la création des graphiques, à la manipulation des tableaux... Tout ce qui concerne directement les logiciels sera traité ici. Vous êtes libres de poser des questions, de poster des captures d'écran, des morceaux de code, des messages d'erreur. Et **vous êtes bien entendus vivement encouragés à vous entraider et à répondre aux questions de vos collègues**. Je n'interviendrai ici que pour répondre aux questions laissées sans réponse ou si les réponses apportées sont inexactes. Le fonctionnement est celui d'un forum de discussion instantané. Vous en tirerez le plus grand bénéfice en participant et en n'ayant pas peur de poser des questions, même si elles vous paraissent idiotes. Rappelez-vous toujours que si vous vous posez une question, d'autres se la posent aussi probablement.

- **#questions-stats** : C'est ici que vous pourrez poser vos questions liées aux méthodes statistiques ou aux choix des modèles de dynamique des populations. Tout ce qui ne concerne pas directement l'utilisation du logiciel (comme par exemple le choix d'un test ou des hypothèses nulles et alternatives, la démarche d'analyse, la signification de tel paramètre ou estimateur, le principe de telle ou telle méthode...) peut être discuté ici. Comme pour le canal **#questions-rstudio**, **vous êtes encouragés à vous entraider et à répondre aux questions de vos collègues**.

Ainsi, quand vous travaillerez à vos TP ou TEA, que vous soyez installés chez vous ou en salle de TP, prenez l'habitude de garder Slack ouvert sur votre ordinateur. Même si vous n'avez pas de question à poser, votre participation active pour répondre à vos collègues est souhaitable et souhaitée. Je vous incite donc fortement à vous **entraider** : c'est très formateur pour celui qui explique, et celui qui rencontre une difficulté a plus de chances de comprendre si c'est quelqu'un d'autre qui lui explique plutôt que la personne qui a rédigé les instructions mal comprises.

Ce document est fait pour vous permettre d'avancer en autonomie et vous ne devriez normalement pas avoir beaucoup besoin de moi si votre lecture est attentive. L'expérience montre en effet que la plupart du temps, il suffit de lire correctement les paragraphes précédents et/ou suivants pour obtenir la réponse à ses questions. J'essaie néanmoins de rester disponible

sur Slack pendant les séances de TP et de TEA de tous les groupes. Cela veut donc dire que même si votre groupe n'est pas en TP, vos questions ont des chances d'être lues et de recevoir des réponses dès que d'autres groupes sont en TP ou TEA. Vous êtes d'ailleurs encouragés à échanger sur Slack aussi pendant vos phases de travail personnel.

Progression conseillée

Si vous avez suivi les livres en ligne des semestres précédents, vous savez que pour apprendre à utiliser **RStudio**, il faut faire les choses soi-même, ne pas avoir peur des messages d'erreurs (il faut d'ailleurs apprendre à les déchiffrer pour comprendre d'où viennent les problèmes), essayer maintes fois, se tromper beaucoup, recommencer, et surtout, ne pas se décourager. J'utilise ce logiciel presque quotidiennement depuis plus de 15 ans et à chaque session de travail, je rencontre des messages d'erreur. Avec suffisamment d'habitude, on apprend à les déchiffrer, et on corrige les problèmes en quelques secondes. Ce livre est conçu pour vous faciliter la tâche, mais ne vous y trompez pas, vous rencontrerez des difficultés, et c'est normal. C'est le prix à payer pour profiter de la puissance du meilleur logiciel permettant d'analyser des données, de produire des graphiques de qualité et de réaliser toutes les statistiques dont vous aurez besoin d'ici la fin de vos études et au-delà.

Pour que cet apprentissage soit le moins problématique possible, il convient de prendre les choses dans l'ordre. C'est la raison pour laquelle les chapitres de ce livre doivent être lus dans l'ordre, et les exercices d'application faits au fur et à mesure de la lecture.

Idéalement, voilà les étapes que vous devriez avoir franchi chaque semaine :

1. La première semaine (09) est consacrée à la comparaison de la moyenne d'une population à une valeur théorique (**?@sec-moy1**). Les fonctions nouvelles sont peu nombreuses, mais ce chapitre est l'occasion d'aborder des notions complexes (test paramétrique ou non, hypothèses nulles et alternatives, p -value, décision, erreurs,

puissance, etc.) qui demanderont une lecture très attentive. C'est également la première fois que vous serez confronté à toutes les étapes d'une analyse de données : de l'importation des données dans **RStudio** jusqu'à la réalisation des tests statistiques et leur interprétation, en passant par le calcul des statistiques descriptives appropriées, la réalisation de graphiques exploratoires pertinents, et la vérification des conditions d'application du test. La maîtrise du logiciel n'est donc ici qu'une petite partie de ce qui est demandé : maîtriser les notions et concepts, comprendre la démarche d'analyse, être capable de l'expliquer et de la reproduire, est ici crucial.

2. La deuxième semaine (10) est consacrée à la comparaison de la moyenne de 2 populations, dans deux situations : lorsque les données des deux groupes sont appariées (**?@sec-moy2**) et quand elles sont indépendantes (**?@sec-moy3**). Ces chapitre seront également l'occasion d'aborder la notion de test unilatéral et bilatéral et d'apprendre comment coder des hypothèses alternatives spécifiques dans **RStudio**. Comme pour la semaine 10, vous aurez l'occasion de mettre en œuvre la totalité de la démarche d'analyse statistique.
3. La troisième semaine (11) sera consacrée à l'analyse de variance (Chapitre 1). Il s'agit du premier exemple de modèle linéaire que nous examinerons ensemble. Ce test est assez différents des autres dans sa philosophie, donc vous aurez besoin d'un peu de temps pour vous approprier la façon de faire, en particulier la logique de la vérification des conditions d'application.
4. La quatrième et dernière semaine (12) sera consacrée à la corrélation (Chapitre 2) et à la régression linéaire (Chapitre 3), le second type de modèle linéaire au programme cette année. Attention, bien que ces chapitres traitent de notions proches, et bien que la façon de traiter la régression soit très proche de celle de l'ANOVA, ces chapitres sont conséquents.

Au final :

- le chapitre 1 doit être traité avant le début de la deuxième séance de TP

- les chapitres 2 et 3 doivent être traités avant le début de la troisième séance de TP
- le chapitre 4 doit être traité avant le début de la quatrième séance de TP
- les chapitres 5 et 6 doivent être traités avant la fin de la quatrième séance de TP
- faute de temps cette année, le chapitre 7, consacré aux tests de comparaison de proportions, ne sera finalement pas traité

Vous comprenez j'espère que dans chaque chapitre, une ou des méthodes vous sont présentées en détail. À la fin de chaque chapitre, un ou des exercices d'application vous sont proposés. À l'issue des ces TP, vous disposerez, dans votre arsenal de biostatisticien, de près d'une quinzaine de tests statistiques différents. Le plus difficile sera d'être en mesure d'identifier lequel choisir face à un jeu de données inconnu, et face à des questions nouvelles. Votre travail consiste donc aussi à vous assurer que vous comprenez bien dans quelle situation utiliser chaque test, et à savoir comment vérifier que vous avez le droit d'utiliser tel ou tel test pour répondre aux questions posées.

Évaluation(s)

L'évaluation de cette partie "TP de Biométrie" de l'EC "Outils pour l'étude et la compréhension du vivant" aura lieu sous la forme d'un devoir à la maison individuel, qui vous demandera de traiter les données acquises dans le cadre du travail de stratégie d'échantillonnage et sur lesquelles vous travaillez depuis plusieurs mois avec Pierrick Bocher. Je vous demanderai de me remettre un script commenté qui me permettra de vérifier les points suivants :

- les grands principes de stratégie d'échantillonnage abordés par Pierrick Bocher sont-ils bien compris ?
- êtes vous capables de choisir la ou les méthodes d'analyses de donnée appropriées pour répondre aux questions posées ?
- êtes vous capables de mettre ces méthode en œuvre dans **RStudio** et d'en interpréter correctement les résultats ?

- maîtrisez-vous suffisamment le logiciel **RStudio** pour réaliser les analyses de données pertinentes (de l'importation des données et leur mise en forme dans le logiciel, à la réalisation et l'interprétation correcte des tests statistiques appropriés, en passant par l'exploration des statistiques descriptives et la création de graphiques informatifs) ?
- enfin, êtes-vous capable de produire un script clair, bien commenté et qui fait ce que vous souhaitez faire sans erreur ?

Pour vous aider à comprendre ce qui est attendu sur ce dernier point, je vous fournis ci-dessous la grille critériée dont je me servirai pour évaluer la forme de votre script. Je ne peux que vous encourager à lire attentivement les critères d'évaluation ci-dessous et à tenter de vous les approprier. Les séances de TP et de TEA qui viennent doivent vous permettre de vous entraîner à produire des scripts de qualité.

Résultat d'apprentissage visé : produire un script clair et fonctionnel permettant d'analyser des données et de communiquer sa démarche d'analyse et ses résultats à ses pairs
Acquis si tous les résultats d'apprentissage sont au moins "Satisfaisants"

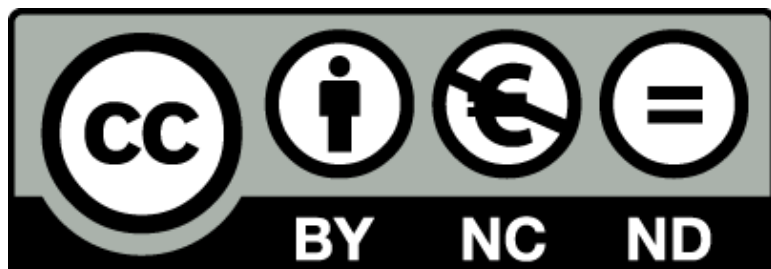
Grille d'évaluation

	Très insuffisant (1 pt)	Insuffisant (2 pts)	Satisfaisant (3 pts)	Très satisfaisant (4 pts)
1. Le script fait ce qu'il est censé faire	Lorsqu'on exécute le script en une fois, de nombreux messages d'erreurs apparaissent, les résultats attendus ne sont pas produits	Lorsqu'on exécute le script en une fois, quelques messages d'erreurs apparaissent, tous les résultats attendus ne sont pas produits, ou certains résultats produits sont faux	Lorsqu'on exécute le script en une fois, aucun message d'erreur n'apparaît, mais certains résultats produits sont faux	Lorsqu'on exécute le script en une fois, aucun message d'erreur n'apparaît, les résultats attendus sont correctement produits
2. Le script est bien commenté	(Presque) pas de commentaires ou commentaires inadaptés	Peu de commentaires sont présents ou s'ils sont présents, la plupart ne permettent pas de comprendre clairement ce qui a été fait par (ou l'intention de) l'auteur du script	La plupart des commandes ou groupes de commandes sont commentés. Certains commentaires manquent parfois de clarté	Chaque commande ou groupe de commande est bien commenté. Les commentaires sont parlants et permettent de comprendre sans ambiguïté l'intention et les choix de l'auteur du script
3. Le script est facile à lire	Le script est difficile à déchiffrer en raison de plusieurs problèmes dans la liste suivante : noms d'objets peu parlants, espaces entre les éléments du code inconsistants ou absents, indentations inexistantes, sauts de lignes manquants ou aux mauvais endroits, code non structuré, ordre des commandes inadéquat	Le script est globalement bien structuré, les commandes apparaissent dans un ordre logique, mais il pourrait être plus lisible car quelques problèmes subsistent dans la liste suivante : noms d'objets peu parlants, espaces entre les éléments du code inconsistants ou absents, indentations inexistantes, sauts de lignes manquants ou aux mauvais endroits	Script bien structuré et lisible. La plupart des éléments de la liste suivante sont respectés : ordre des commandes adéquat, noms d'objets parlants, espaces, indentations et sauts de lignes consistants, bon équilibre entre commentaires et commandes R	Script parfaitement structuré et lisible. Tous les éléments de syntaxe sont respectés (espaces, ponctuation, indentation, sauts de lignes...), les noms d'objets sont courts et parlants, les commandes sont correctement ordonnées et un bon équilibre est respecté entre code et commentaires
4. Le script est générique	Non. Si les données changent (ajout ou suppression de lignes dans le tableau de départ, changements de valeurs...), le script ne fonctionne plus (des messages d'erreurs apparaissent ou les résultats produits sont faux) - 1pt	Oui. Si les données changent (ajout ou suppression de lignes dans le tableau de départ, changements de valeurs...), le script fonctionne toujours : il ne renvoie pas de message d'erreur et les résultats fournis reflètent les modifications des données - 2 pts		

L'énoncé de ce devoir sera déposé sur Moodle le samedi 25 mars, et vos travaux seront à rendre avant votre départ en stage, le samedi 15 avril à midi au plus tard.

Licence

Ce livre est ligne est sous licence Creative Commons ([CC BY-NC-ND 4.0](#))



Vous êtes autorisé à partager, copier, distribuer et communiquer ce matériel par tous moyens et sous tous formats, tant que les conditions suivantes sont respectées :

① **Attribution** : vous devez créditer ce travail (donc citer son auteur), fournir un lien vers ce livre en ligne, intégrer un lien vers la licence Creative Commons et indiquer si des modifications du contenu original ont été effectuées. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l’auteur vous soutient ou soutient la façon dont vous avez utilisé son travail.

Ⓢ **Pas d’Utilisation Commerciale** : vous n’êtes pas autorisé à faire un usage commercial de cet ouvrage, ni de tout ou partie du matériel le composant. Cela comprend évidemment la diffusion sur des plateformes de partage telles que studocu.com qui tirent profit d’œuvres dont elles ne sont pas propriétaires, souvent à l’insu des auteurs.

⊖ **Pas de modifications** : dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l’ouvrage original, vous n’êtes pas autorisé à distribuer ou mettre à disposition l’ouvrage modifié.

🔒 **Pas de restrictions complémentaires** : vous n’êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser cet ouvrage dans les conditions décrites par la licence.

1 Comparaison de moyennes : plus de 2 groupes

1.1 Pré-requis

Comme pour chaque nouveau chapitre, je vous conseille de travailler dans un nouveau script que vous placerez dans votre répertoire de travail, et dans une nouvelle session de travail (Menu **Session** > **Restart R**). Inutile en revanche de créer un nouveau **Rproject** : vous pouvez tout à fait avoir plusieurs script dans le même répertoire de travail et pour un même **Rproject**. Comme toujours, consultez [le livre en ligne du semestre 3](#) si vous ne savez plus comment faire.

Si vous êtes dans une nouvelle session de travail (ou que vous avez quitté puis relancé **RStudio**), vous devrez penser à recharger en mémoire les packages utiles. Dans ce chapitre, vous aurez besoin d'utiliser :

- le **tidyverse** (Wickham 2023), qui comprend notamment le package **readr** (Wickham, Hester, et Bryan 2023), pour importer facilement des fichiers **.csv** au format **tibble**, le package **dplyr** (Wickham, François, et al. 2023), pour manipuler des tableaux, et le package **ggplot2** (Wickham, Chang, et al. 2023) pour les représentations graphiques.
- **readxl** (Wickham et Bryan 2023), pour importer facilement des fichiers Excel au format **tibble**.
- **skimr** (Waring et al. 2022), qui permet de calculer des résumés de données très informatifs.
- **car** (Fox, Weisberg, et Price 2023), qui permet d'effectuer le test de comparaison des variances de Levene.
- **broom** (Robinson, Hayes, et Couch 2023), qui fait partie du **tidyverse** mais qu'il faut charger explicitement. La fonction **tidy()** de ce package nous permettra de "ranger" correctement les résultats de tests dans un **tibble**.

- `DescTools` (Signorell 2023), afin de réaliser un test spécifique de comparaisons multiples. N’oubliez pas de l’installer si nécessaire, avant de le charger en mémoire.
- le package `palmerpenguins` (Horst, Hill, et Gorman 2022) pour accéder au jeu de données `penguins` que nous utiliserons pour les exercices d’application.

```
library(tidyverse)
library(readxl)
library(skimr)
library(car)
library(broom)
library(DescTools)
library(palmerpenguins)
```

Vous aurez également besoin des jeux de données suivants que vous pouvez dès maintenant télécharger dans votre répertoire de travail :

- [Light.csv](#)
- [Insectes.csv](#)

Enfin, je spécifie ici une fois pour toutes le thème que j’utiliserai pour tous les graphiques de ce chapitre. Libre à vous de choisir un thème différent ou de vous contenter du thème proposé par défaut :

```
theme_set(theme_bw())
```

1.2 Contexte

Voyager dans un pays éloigné peut faire souffrir de décalage horaire. Habituellement, la resynchronisation de l’horloge interne circadienne dans le nouveau fuseau horaire est réalisée grâce à la perception de la lumière par les yeux. Ce changement progressif du rythme de notre horloge interne est appelé “décalage de phase”. Ce phénomène a été étudié par 2 chercheurs en 1998 (Campbell et Murphy 1998), qui ont montré que ce décalage de phase pouvait également être obtenu en exposant des sujets à la lumière, non pas au niveau de leurs

yeux, mais au niveau de leur fosse (ou creux) poplitée, c'est-à-dire, derrière les genoux.

Cette découverte a été vivement critiquée par certains, et saluée comme une découverte majeure par d'autres. Toutefois, certains aspects du design expérimental de l'étude de 1998 ont été mis en doute en 2002 : il semble en effet que lors de l'exposition du creux poplité, les yeux de certains patients ont été également exposés à de faibles intensités lumineuses. Pour vérifier les trouvailles de Campbell et Murphy, Wright et Czeisler (Wright et Czeisler 2002) ont ré-examiné ce phénomène. La nouvelle expérience a évalué les rythmes circadiens en mesurant les cycles quotidiens de production de mélatonine chez 22 participants placés au hasard dans 3 groupes. Les patients étaient réveillés en pleine nuit et exposés :

1. Soit à 3 heures de lumière appliquée exclusivement derrière leurs genoux (groupe `knee`).
2. Soit à 3 heures de lumière appliquée exclusivement à leurs yeux (groupe `eyes`).
3. Soit à 3 heures d'obscurité totale (groupe `control`).

Le décalage de phase du cycle de production de mélatonine était mesuré 48h plus tard. Des chiffres négatifs indiquent un retard de production de mélatonine. C'est l'effet théorique attendu du traitement lumineux administré. Un décalage de phase positif indique une production de mélatonine plus précoce. Une absence de changement se traduit par un décalage de phase de 0.

1.3 Importation et mise en forme des données

Les données brutes de cette étude sont fournies dans le fichier [Light.csv](#). Importez ces données dans RStudio et examinez les données brutes grâce à la fonction `View()`.

```
Light
```

```
# A tibble: 22 x 2
  treatment shift
  <chr>         <dbl>
```

```

1 control    0.53
2 control    0.36
3 control    0.2
4 control   -0.37
5 control   -0.6
6 control   -0.64
7 control   -0.68
8 control   -1.27
9 knee       0.73
10 knee      0.31
# i 12 more rows

```

Le tableau obtenu est-il au format long ou au format court/large ? Pourquoi un tableau au format suivant n'aurait-il pas de sens ?

```

# A tibble: 8 x 3
  control eyes knee
  <dbl> <dbl> <dbl>
1    0.53 -0.78  0.73
2    0.36 -0.86  0.31
3    0.2  -1.35  0.03
4   -0.37 -1.48 -0.29
5   -0.6  -1.52 -0.56
6   -0.64 -2.04 -0.96
7   -0.68 -2.83 -1.61
8   -1.27 NA    NA

```

Lorsque l'on réalise une analyse de variance, puisque les effectifs ne sont pas nécessairement identiques dans tous les groupes (c'est ce qu'on appelle un design déséquilibré, ou "unbalanced design"), présenter les tableaux au format long est indispensable. Par ailleurs, notez que les ANOVAs réalisées sur des "balanced design" (ou designs équilibrés, pour lesquels tous les groupes sont de même taille), sont beaucoup plus puissantes que les ANOVAs réalisées sur des "unbalanced designs".

Ici, le tableau de données est très simple (et de petite taille). Il n'y a pas de données manquantes et aucune création de nouvelle variable n'est nécessaire. La seule modification que nous devrions faire est de transformer la variable **treatment** en facteur :

```
Light <- Light %>%
  mutate(treatment = factor(treatment))
```

Comme toujours, les niveaux du facteur sont automatiquement classés par ordre alphabétique :

```
levels(Light$treatment)
```

```
[1] "control" "eyes"    "knee"
```

Pour les statistiques descriptives et les graphiques qui viendront après, nous souhaitons indiquer l'ordre suivant : control, puis knee, puis eyes :

```
Light <- Light %>%
  mutate(treatment = fct_relevel(treatment, "control", "knee", "eyes"))
```

```
Light
```

```
# A tibble: 22 x 2
  treatment shift
  <fct>      <dbl>
1 control    0.53
2 control    0.36
3 control    0.2
4 control   -0.37
5 control   -0.6
6 control   -0.64
7 control   -0.68
8 control   -1.27
9 knee       0.73
10 knee      0.31
# i 12 more rows
```

```
Light$treatment
```

```
[1] control control control control control control control control knee
[10] knee     knee     knee     knee     knee     knee     eyes     eyes     eyes
[19] eyes     eyes     eyes     eyes
Levels: control knee eyes
```

Attention à bien respecter la casse (le respect des majuscules/minuscules est toujours aussi important dans RStudio).

1.4 Exploration statistique des données

Comme toujours, et maintenant que nos données sont au bon format, il est nécessaire d'examiner quelques statistiques descriptives pour chaque catégorie étudiée. On peut tout d'abord commencer par examiner la taille de chaque échantillon :

```
Light %>%  
  count(treatment)
```

```
# A tibble: 3 x 2  
  treatment     n  
  <fct>      <int>  
1 control      8  
2 knee         7  
3 eyes         7
```

Nous avons ici la confirmation que le design expérimental n'est pas équilibré, puisque le groupe **control** compte un individu de plus. Nous pouvons ensuite utiliser la fonction **skim** du package **skimr** pour obtenir un résumé des données :

```
Light %>%  
  group_by(treatment) %>%  
  skim()
```

```
-- Data Summary -----  
  
Name                               Values  
Number of rows                     Piped data  
Number of rows                     22  
Number of columns                   2  
  
-----  
Column type frequency:  
  numeric                           1  
-----
```

Group variables

treatment

```
-- Variable type: numeric -----
  skim_variable treatment n_missing complete_rate  mean    sd    p0    p25
1 shift          control         0             1 -0.309 0.618 -1.27 -0.65
2 shift          knee           0             1 -0.336 0.791 -1.61 -0.76
3 shift          eyes           0             1 -1.55  0.706 -2.83 -1.78
      p50    p75  p100 hist
1 -0.485  0.24  0.53
2 -0.29   0.17  0.73
3 -1.48  -1.10 -0.78
```

Il semble que le groupe **eyes** se comporte un peu différemment des autres groupes. En effet, pour les groupes **control** et **knee**, les valeurs des indices de position observés sont très proches :

- les moyennes et les médianes sont négatives mais proches de 0.
- les valeurs observées sont négatives pour certaines, et positives pour d'autres (la colonne **p0** contient les minimas et la colonne **p100** contient les maximas).

En revanche, pour le groupe **eyes**, les décalages de phase observés sont tous négatifs (le maximum, présenté dans la colonne **p100** vaut -0.78) et la moyenne est près de 5 fois plus faible que pour les 2 autres groupes.

Concernant la dispersion, les écart-types semblent en revanche très proches dans les 3 groupes (entre 0.6 et 0.8).

Enfin, les histogrammes présentés pour chaque groupe semblent très éloignés d'une distribution Normale. C'est logique compte tenu des faibles effectifs dans chaque groupe. Nous verrons plus tard que cela n'a aucune importance puisque les conditions d'application de l'ANOVA portent sur les **résidus de l'ANOVA** (nous verrons plus loin de quoi il s'agit), et pas sur les données brutes.

Il semble donc que seul le groupe **eyes** soit véritablement différent du groupe témoin. Pour le vérifier, on peut calculer les intervalles de confiance à 95% des moyennes. Nous examinerons ensuite quelques graphiques, puis nous ferons un test d'hypothèses.

```
Light %>%
  reframe(mean_cl_normal(shift), .by = treatment)
```

```
# A tibble: 3 x 4
  treatment      y   ymin   ymax
  <fct>      <dbl> <dbl> <dbl>
1 control  -0.309 -0.825  0.208
2 knee    -0.336 -1.07   0.396
3 eyes    -1.55  -2.20  -0.898
```

Là encore, le groupe **eyes** semble assez différent des 2 autres. L'intervalle de confiance à 95% de ce groupe ([-2.20 ; -0.90]) est totalement disjoint du groupe **knee** ([-1.07 ; 0.40]) et chevauche à peine celui du groupe **control** ([-0.82 ; 0.21]). L'intervalle de confiance du groupe **knee** recouvre en revanche en totalité celui du groupe **control**. Il n'y aura donc vraisemblablement pas de différence significative entre ces 2 groupes, mais une différence significative entre le groupe **eyes** et les 2 autres. Pour visualiser un peu mieux ces résultats préliminaires, examinons quelques graphiques.

1.5 Exploration graphique

Comme toujours, il est indispensable de regarder à quoi ressemblent les données brutes sur un ou des graphiques. Les statistiques descriptives ne racontent en effet pas toujours toute l'histoire. Ici, nous allons superposer les données brutes, sous forme de nuage de points, aux boîtes à moustaches :

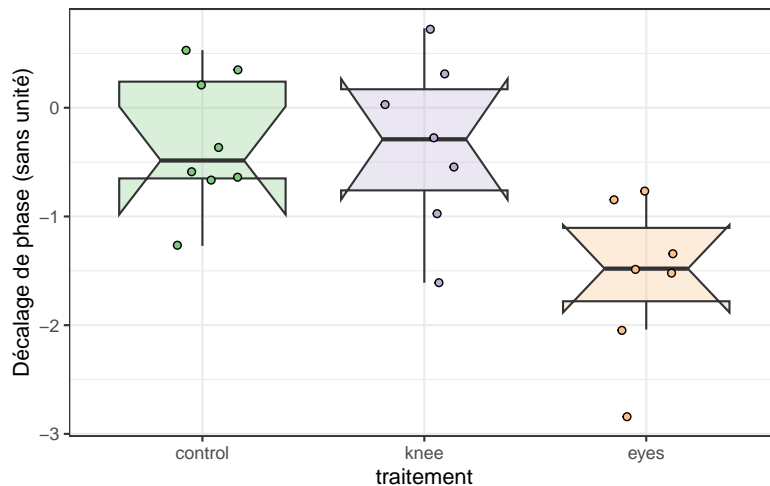
```
Light %>%
  ggplot(aes(x = treatment, y = shift, fill = treatment)) +
  geom_boxplot(notch = TRUE, show.legend = FALSE,
              alpha = 0.3, outlier.colour = NA) +
  geom_jitter(width = 0.2, shape = 21, show.legend = FALSE) +
  labs(x = "traitement", y = "Décalage de phase (sans unité)") +
  scale_fill_brewer(palette = "Accent")
```

```
Notch went outside hinges
i Do you want `notch = FALSE`?
```

```

Notch went outside hinges
i Do you want `notch = FALSE`?
Notch went outside hinges
i Do you want `notch = FALSE`?

```



Puisqu'il y a peu de données, les intervalles de confiance à 95% sont très larges. Ils dépassent d'ailleurs presque systématiquement les quartiles, ce qui explique l'apparence bizarre des boîtes à moustaches et les messages d'avertissement affichés lors de la création du graphique. Il vaudrait donc mieux représenter cette figure sans ces intervalles de confiance. Toutefois, avant de les retirer, on peut constater ici que les IC 95% se chevauchent complètement pour les séries **control** et **knee**. En revanche, il n'y a aucun chevauchement de l'IC 95% du groupe **eyes** avec les 2 autres groupes. Ces résultats sont très légèrement différents de ceux obtenus plus haut, car on examine ici les intervalles de confiance à 95% des médianes, alors qu'on regardait les intervalles de confiance à 95% des moyennes dans la section précédente. Les conclusions sont toutefois les mêmes : on s'attend donc à trouver une différence de moyenne significative entre le groupe **eyes** d'une part, et les groupes **control** et **knee** d'autre part, mais pas de différence de moyenne entre les groupes **control** et **knee**.

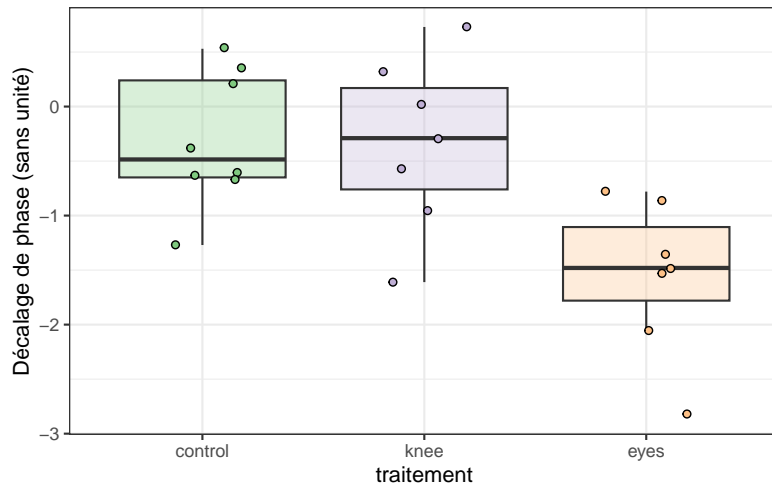
```

Light %>%
  ggplot(aes(x = treatment, y = shift, fill = treatment)) +
  geom_boxplot(show.legend = FALSE,
               alpha = 0.3, outlier.colour = NA) +

```



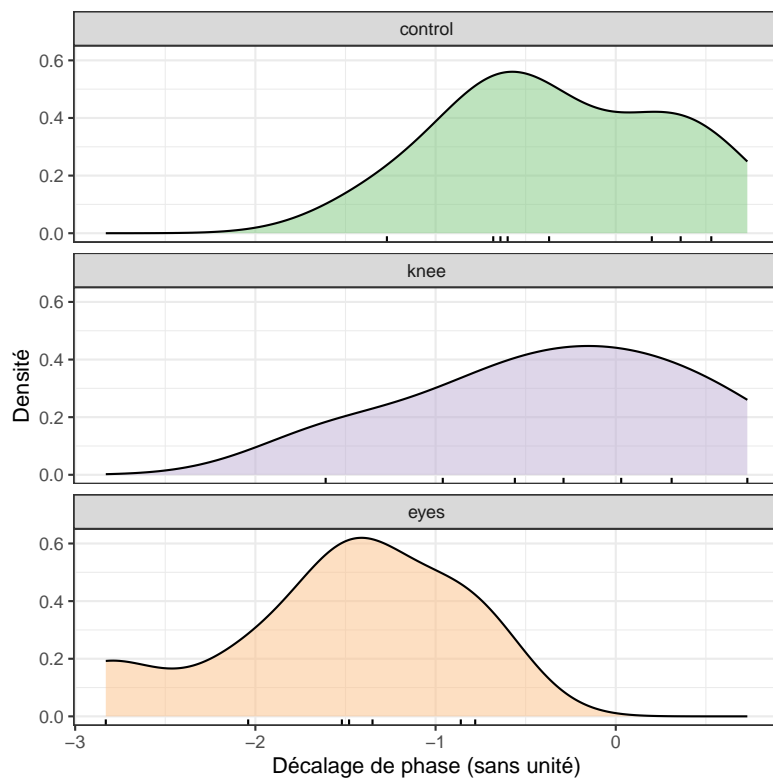
```
geom_jitter(width = 0.2, shape = 21, show.legend = FALSE) +
labs(x = "traitement", y = "Décalage de phase (sans unité)") +
scale_fill_brewer(palette = "Accent")
```



On constate ici visuellement que les 3 séries ont une étendue à peu près similaire, et que le groupe **eyes** semble se distinguer des 2 autres par des valeurs plus faibles. Enfin, les boîtes contenant 50% des valeurs centrales (donc l'étendue des valeurs entre les premiers et troisièmes quartiles) recouvrent le 0 pour les 2 groupes **control** et **knee**, mais pas pour **eyes**.

L'examen d'un graphique de densité facetté donne les mêmes informations :

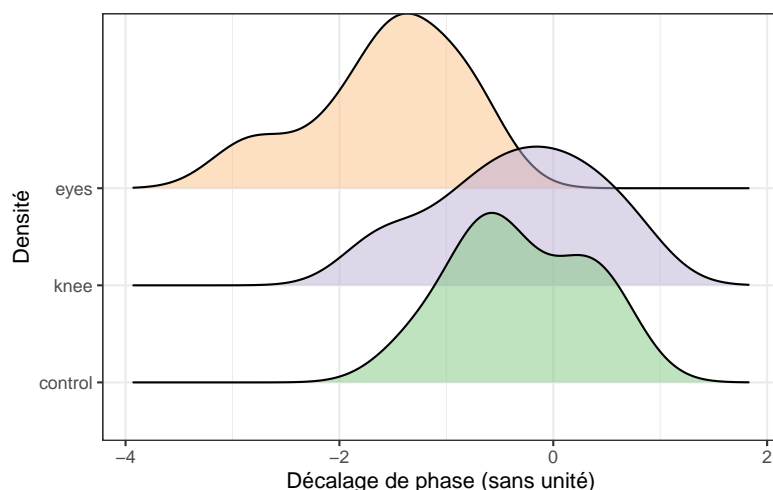
```
Light %>%
  ggplot(aes(x = shift, fill = treatment)) +
  geom_density(show.legend = FALSE, alpha = 0.5) +
  geom_rug() +
  facet_wrap(~treatment, ncol = 1) +
  scale_fill_brewer(palette = "Accent") +
  labs(x = "Décalage de phase (sans unité)", y = "Densité")
```



Un package utile lorsque l'on dispose d'un grand nombre de groupes que l'on souhaite comparer à l'aide de graphiques de densité est le package `ggridges` :

```
library(ggridges)
Light %>%
  ggplot(aes(x = shift, y = treatment, fill = treatment)) +
  geom_density_ridges(show.legend = FALSE, alpha = 0.5) +
  scale_fill_brewer(palette = "Accent") +
  labs(x = "Décalage de phase (sans unité)", y = "Densité")
```

Picking joint bandwidth of 0.366



1.6 Le test paramétrique

Le test paramétrique permettant de comparer la moyenne de plusieurs populations en une seule étape est l'**analyse de variance à un facteur**. Contrairement aux tests que nous avons vus jusqu'à maintenant, les conditions d'application de ce test ne seront vérifiées qu'**après** avoir réalisé l'analyse. En effet, les conditions d'application de l'ANOVA ne se vérifient pas sur les données brutes mais sur les **résidus de l'ANOVA**. C'est d'ailleurs ce que l'on appelle l'**analyse des résidus**, ou diagnostic de l'ANOVA.

1.6.1 Réalisation du test

Dans R, l'analyse de variance se fait grâce à la fonction `aov()` (comme "Analysis Of Variance"). La syntaxe est la même que pour un certain nombre de tests déjà vus dans les chapitres précédents : il faut fournir une formule à la fonction. On place la variable numérique expliquée à gauche du `~`, et à droite, la variable qualitative explicative (le facteur).

Contrairement aux autres tests réalisés jusqu'ici, les résultats du test devront être sauvegardés dans un objet. Outre les résultats du test, cet objet contiendra également tous les éléments permettant de vérifier si les conditions d'application de l'ANOVA sont réunies ou non.

Les hypothèses testées sont les suivantes :

- H_0 : les moyennes de toutes les populations sont égales ($\mu_{\text{control}} = \mu_{\text{knee}} = \mu_{\text{eyes}}$).
- H_1 : toutes les moyennes ne sont pas égales. Au moins l'une d'entre elles diffère des autres.

```
# Réalisation de l'ANOVA 1 facteur
res <- aov(shift ~ treatment, data = Light)

# Affichage des résultats
res
```

Call:

```
aov(formula = shift ~ treatment, data = Light)
```

Terms:

	treatment	Residuals
Sum of Squares	7.224492	9.415345
Deg. of Freedom	2	19

Residual standard error: 0.7039492

Estimated effects may be unbalanced

L'affichage des résultats bruts ne nous apprend que peu de choses. En revanche, la fonction `summary()` donne la réponse du test :

```
summary(res)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	2	7.224	3.612	7.289	0.00447 **
Residuals	19	9.415	0.496		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pour le facteur étudié (`treatment`), on obtient le nombre de degrés de libertés (`Df`), la somme des carrés (`Sum Sq`), les carrés moyens (`Mean Sq`), la statistique du test (`F`) et la p -value (`Pr(>F)`). Ici, la p -value est inférieure à α , donc on rejette H_0 . Au moins l'une des moyennes est différente des autres.

Avant d'aller plus loin dans l'interprétation de ces résultats, il nous faut déterminer si nous avons bel et bien le droit de réaliser cette ANOVA, en vérifiant si les conditions d'application de l'ANOVA sont remplies.

1.6.2 Conditions d'application

L'ANOVA est un test paramétrique, et comme pour tous les tests paramétriques, des conditions d'application doivent être vérifiées pour avoir le droit d'effectuer le test. À la différence des autres tests paramétriques que nous avons réalisés jusqu'ici, les conditions d'application de l'ANOVA ne doivent pas être vérifiées **avant** de faire le test, mais **après**.

Les résultats de l'ANOVA ne seront donc valides que si les conditions d'application sont vérifiées. Comme indiqué plus haut, ces conditions d'application doivent être vérifiées sur **les résidus de l'ANOVA**, donc nécessairement **après** avoir réalisé l'analyse. Les résidus de l'ANOVA représentent l'écart entre chaque observation et la moyenne de son groupe, et ils sont calculés au moment où nous réalisons l'ANOVA.

! Important

Les conditions d'application de l'ANOVA ne se vérifient pas sur les données brutes comme c'est le cas du test de Student, mais sur **les résidus de l'ANOVA**, qui sont calculés au moment où l'ANOVA est réalisée. Par conséquent, on ne vérifie pas les conditions d'application avant mais bien **après** avoir fait le test.

Ça n'est que si les conditions d'application sont remplies qu'on aura le droit d'interpréter les résultats de l'ANOVA.

Pour que l'ANOVA soit valide, les résidus doivent :

1. Être indépendants.
2. Être homogènes.
3. Être distribués normalement.

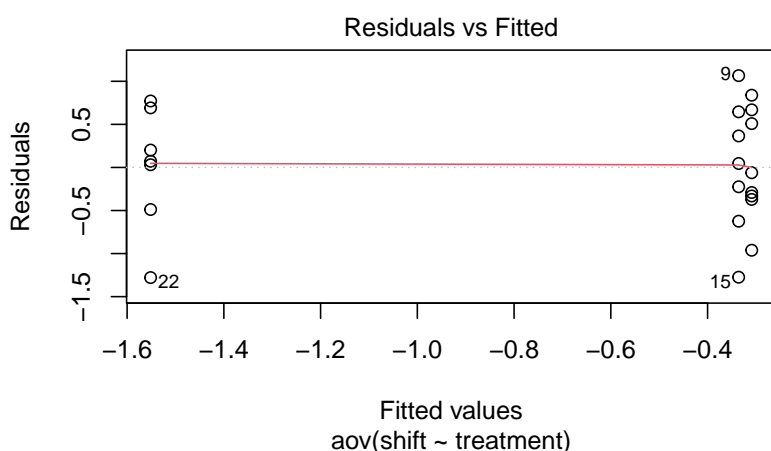
1.6.2.1 Indépendance des résidus

L'indépendance des résidus signifie que connaître la valeur d'un résidu ne permet pas de prédire la valeur d'un autre résidu. Si les données ont été collectées correctement (échantillonnage aléatoire simple, indépendance des observations), on considère généralement que cette condition est vérifiée. Les 2 autres conditions d'application se vérifient soit graphiquement, soit avec un test d'hypothèses.

1.6.2.2 Homogénéité des résidus

L'homogénéité des résidus signifie que les résidus doivent avoir à peu près la même variance pour chacun des groupes comparés. On peut vérifier que cette condition d'application est vérifiée grâce à ce graphique, qui représente les résidus (**residuals**, sur l'axe des y) en fonction des valeurs ajustées (c'est-à-dire la moyenne de chaque groupe, **fitted values**, sur l'axe des x) :

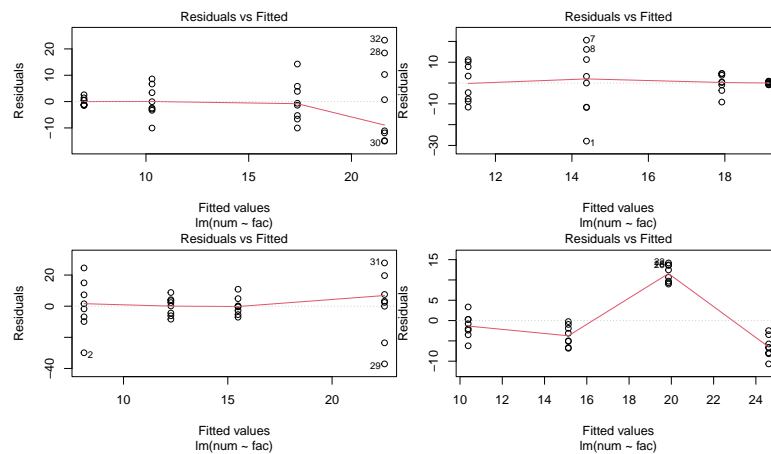
```
plot(res, which = 1)
```



Ici, les résidus sont considérés comme homogènes car nous avons à peu près autant de résidus positifs que négatifs et que la ligne rouge est très proche du 0. L'étalement (vertical) des résidus est à peu près le même de la gauche à la droite du graphique. On pourrait donc faire entrer les résidus dans

une boîte rectangulaire horizontale centrée sur le 0 et ayant la même largeur d'un bout à l'autre du graphique.

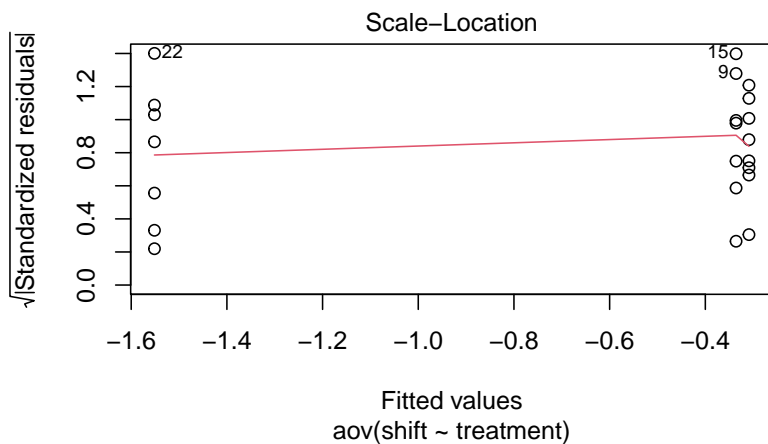
Ci-dessous, j'affiche quelques exemples de situations où les résidus ne sont pas homogènes afin que vous puissiez voir à quoi ressemblent les graphiques des résidus en fonction des valeurs ajustées dans ce type de situation :



Pour ces 4 graphiques, les résidus ne rentrent pas dans une boîte rectangulaire qui a la même hauteur d'un bout à l'autre du graphique. Les résidus de ces ANOVAs fictives ne sont donc pas homogènes et les conditions d'application de l'ANOVA ne sont donc pas réunies.

Mais revenons à nos données de décalage de phase. Une autre façon de visualiser les résidus est d'utiliser le graphique suivant :

```
plot(res, which = 3)
```



Sur ce graphique, ce qui compte principalement, c'est la droite en rouge. Elle est ici presque horizontale, ce qui montre que les résidus de tous les groupes (un groupe à gauche et 2 à droite) ont à peu près même moyenne.

Enfin, cette condition d'homogénéité des résidus entre les groupes peut également être vérifiée grâce au test de Levene. Pour ce test, les hypothèses seront les suivantes :

- H_0 : les résidus sont homogènes (*i.e.* identiques dans tous les groupes).
- H_1 : les résidus ne sont pas homogènes (*i.e.* au moins un groupe présente des résidus dont la variance est différente des autres).

```
leveneTest(res$residuals ~ Light$treatment)
```

```
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 2  0.1586 0.8545
    19
```

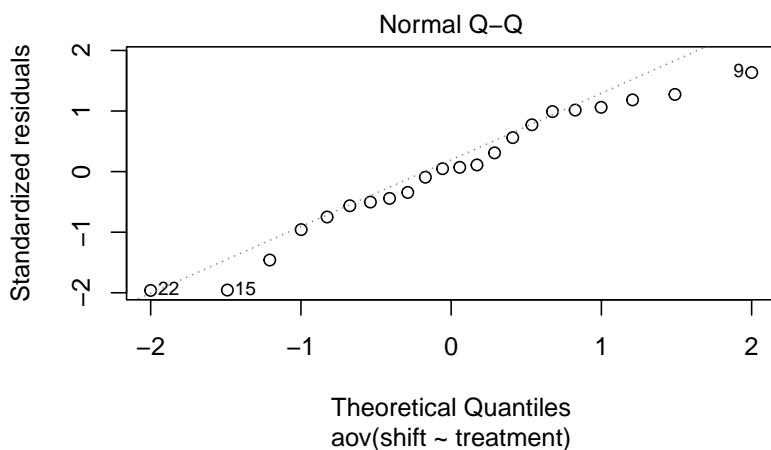
Ici, puisque $p > \alpha$, on ne peut pas rejeter l'hypothèse nulle. Les résidus sont donc bien homogènes.

Reste à vérifier la normalité des résidus.

1.6.2.3 Normalité des résidus

Comme pour l'homogénéité des résidus, leur normalité peut être examinée graphiquement ou avec un test statistique :

```
plot(res, which = 2)
```



Sur un graphique quantile-quantile comme celui-là, on considère que les observations sont distribuées normalement si les points sont bien alignés sur la droite. Ici, la plupart des points sont très proches de la droite, ce qui laisse penser que les résidus suivent bien la loi Normale. Mais il est souvent difficile, surtout pour les néophytes, de savoir à partir de quel écart entre les points et la droite il faut considérer que les résidus n'ont pas une distribution Normale.

On peut donc confirmer (ou non !) notre première impression avec le test de Normalité de Shapiro-Wilk. Il s'agit du même test de Normalité que nous utilisons sur les données brutes pour vérifier les conditions d'application du test de Student. Ici, on applique ce test sur les résidus de l'ANOVA :

```
shapiro.test(res$residuals)
```

Shapiro-Wilk normality test

```
data: res$residuals  
W = 0.95893, p-value = 0.468
```

Comme pour tous les tests de Shapiro-Wilk, l'hypothèse nulle est la normalité des observations. Ici, puisque $p > \alpha$, on ne peut pas rejeter H_0 . Les résidus suivent donc bien la loi Normale.

Toutes les conditions d'application de l'ANOVA sont donc vérifiées. Nous avons donc bien le droit de la réaliser et ses résultats sont valides.

Dernière chose, il est possible de produire les 3 graphiques ci-dessus (et même un quatrième que nous ne décrivons pas ici), en une seule commande :

```
plot(res)
```

Il faut alors presser la touche **Entrée** de votre clavier pour afficher successivement les 4 graphiques produits.

1.6.3 Interprétation des résultats

Maintenant que nous avons la confirmation que les conditions d'application sont vérifiées, revenons aux résultats de l'ANOVA :

```
summary(res)
```

```

              Df Sum Sq Mean Sq F value    Pr(>F)
treatment      2  7.224   3.612    7.289 0.00447 **
Residuals     19  9.415   0.496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comme indiqué plus haut, la première ligne du tableau d'ANOVA contient toutes les informations pertinentes pour interpréter ces résultats. En particulier, la dernière valeur de la ligne correspond à la p -value. Ici, elle est inférieure au seuil α de 0.05. On pourrait rédiger les résultats de cette analyse ainsi :

Une analyse de variance montre que la moyenne des 3 groupes n'est pas identique ($F = 7.289$, $p = 0.004$). Un test de Levene a permis de vérifier la condition d'homogénéité de la variance des résidus ($F = 0.189$, $p = 0.855$), et un test de Shapiro-Wilk a confirmé la normalité des résidus ($W = 0.959$, $p = 0.468$).

Ainsi, puisque $p < \alpha$, on rejette H_0 . ON a donc bien montré que tous les groupes n'avaient pas la même moyenne. Mais à ce stade, on ne sait pas encore si tous les groupes ont des moyennes strictement différentes les unes des autres, ou si seul un groupe (et lequel) présente une moyenne différente des 2 autres.

À l'issue de cette analyse, deux questions restent donc en suspens :

1. Entre quels groupes les moyennes sont-elles significativement différentes ?
2. Quelle est la magnitude de ces différences ?

Pour répondre à ces 2 questions, il nous faut réaliser des tests *a posteriori* ou tests *post-hoc*.

Les tests *post-hoc*

Les tests *post-hoc* doivent être réalisés uniquement si l'hypothèse nulle de l'ANOVA est rejetée. Ils sont alors nécessaires pour déterminer entre quels groupes les moyennes sont significativement différentes. Si à l'inverse, l'ANOVA n'a pas permis de rejeter H_0 , alors on peut conclure à l'absence de différence de moyenne entre les groupes (*i.e.* toutes les moyennes sont égales), et les tests *post-hoc* n'ont alors aucun intérêt.

1.6.4 Tests *a posteriori* ou tests *post-hoc*

Lorsqu'une ANOVA montre que tous les groupes n'ont pas la même moyenne, il faut en théorie effectuer toutes les comparaisons de moyennes deux à deux possibles. Le problème est que lorsque l'on effectue des comparaisons multiples, les erreurs α (probabilité de rejeter à tort H_0) de tous les tests s'ajoutent. Ainsi :

- pour comparer 3 groupes 2 à 2, nous avons besoin de 3 tests.
- Pour comparer 4 groupes 2 à 2, nous avons besoin de 6 tests.
- pour comparer 5 groupes 2 à 2, nous avons besoin de 10 tests.
- pour comparer 6 groupes 2 à 2, nous avons besoin de 15 tests.
- pour comparer k groupes 2 à 2, nous avons besoin de $\frac{k(k-1)}{2}$ tests.

Ici, puisque pour chaque test, un risque α de 5% de rejeter à tort l'hypothèse nulle est commis, réaliser 3 tests ferait monter le risque de s'être trompé quelque part à 15%. C'est la raison pour laquelle des tests spécifiques existent. Nous en verrons 2 : le test de comparaisons multiples de Student et le test de Tukey (ou "Honestly Significant Difference Test"). Pour ces tests, des précautions sont prises qui garantissent que le risque α **global** (à l'issue de l'ensemble des tests) est maîtrisé et qu'il reste fixé à 5%, quel que soit le nombre de comparaisons effectuées.

1.6.4.1 Comparaisons multiples de Student

Le test de comparaisons multiples de Student est réalisé avec la fonction `pairwise.t.test()`. En réalité, ici, 3 tests de Student seront réalisés. Les p -values des tests seront simplement modifiées afin que globalement, le risque α n'augmente pas. Pour chaque test réalisé, les hypothèses nulles et alternatives sont les mêmes que celles décrites à la ?@sec-student :

- H_0 : la moyenne des deux populations est égale ($\mu_1 = \mu_2$, soit $\mu_1 - \mu_2 = 0$).
- H_1 : la moyenne des deux populations est différente ($\mu_1 \neq \mu_2$, soit $\mu_1 - \mu_2 \neq 0$).

Attention, pour ce test, la syntaxe "formules", qui utilise le tilde (\sim) n'est pas possible. Il faut obligatoirement fournir à la fonction 2 objets : la colonne contenant la variable expliquée numérique, et la colonne (facteur) contenant les catégories (ici, le facteur contenant le type de traitement appliqué à chaque individu lors de l'expérience) :

```
# Réalisation du test
post_hoc1 <- pairwise.t.test(Light$shift, Light$treatment)

# affichage des résultats
post_hoc1
```

Pairwise comparisons using t tests with pooled SD

data: Light\$shift and Light\$treatment

```
      control knee
knee 0.9418  -
eyes 0.0088  0.0088
```

P value adjustment method: holm

Seules les p -values de chaque test sont fournies sous la forme d'une demi-matrice. On constate ainsi qu'une seule p -value est supérieure à $\alpha = 0.05$: celle du test comparant les moyennes des groupes `knee` et `control`. Une autre façon de visualiser ces résultats consiste à utiliser la fonction `tidy()` du package `broom` que nous avons mis en mémoire un peu plus tôt. Les résultats seront les mêmes. Ils seront simplement rangés dans un `tibble` :

```
tidy(post_hoc1)
```

```
# A tibble: 3 x 3
  group1 group2 p.value
  <chr>   <chr>   <dbl>
1 knee   control 0.942
2 eyes   control 0.00879
3 eyes   knee    0.00880
```

Nous avons donc la confirmation que les moyennes des groupes `knee` et `control` ne sont pas significativement différentes l'une de l'autre. En revanche, la moyenne du groupe `eyes` est différente de celle des 2 autres groupes ($p = 0.009$ pour les 2 tests).

Nous avons donc appris des choses nouvelles, mais nous ne savons toujours pas quelle est la magnitude de la différence détectée entre le groupe `eyes` et les 2 autres. Le test de Tukey HSD nous permet de répondre à cette question.

1.6.4.2 Test de Tukey

Ce test est souvent plus avantageux que le test des comparaisons multiples de Student, car outre la p -value de chaque comparaison deux à deux, il renvoie des informations concernant les différences de moyennes entre chaque paire de modalités du facteur étudié, et les intervalles de confiance à 95% de ces différences de moyennes. Donc en plus de savoir quels groupes ou traitements sont significativement différents les uns des autres, ce test nous indique l'importance des différences détectées.

Pour effectuer ce test, on utilise la fonction `TukeyHSD()`, à laquelle on fournit simplement l'objet contenant les résultats de l'ANOVA :

```
# Réalisation du test de Tukey HSD
post_hoc2 <- TukeyHSD(res)

# Affichage des résultats
post_hoc2
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = shift ~ treatment, data = Light)
```

```
$treatment
```

	diff	lwr	upr	p adj
knee-control	-0.02696429	-0.9525222	0.8985936	0.9969851
eyes-control	-1.24267857	-2.1682364	-0.3171207	0.0078656
eyes-knee	-1.21571429	-2.1716263	-0.2598022	0.0116776

Nous obtenons bien à la fois la p -value des comparaisons 2 à 2, ainsi que l'estimation des différences de moyennes (et de leur intervalle de confiance à 95%) entre paires de groupes. Là encore, l'utilisation de la fonction `tidy()` du package `broom`

peut rendre les résultats plus lisibles (ou en tous cas, plus faciles à manipuler) :

```
tidy(post_hoc2)
```

```
# A tibble: 3 x 7
```

	term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	treatment	knee-control	0	-0.0270	-0.953	0.899	0.997
2	treatment	eyes-control	0	-1.24	-2.17	-0.317	0.00787
3	treatment	eyes-knee	0	-1.22	-2.17	-0.260	0.0117

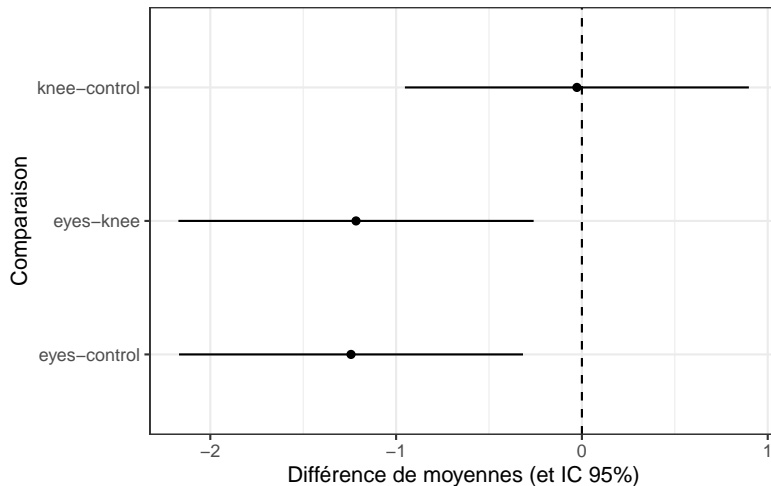
La première ligne de ce tableau nous confirme une absence de différence de moyenne significative entre les groupes **knee** et **control** ($p = 0.997$). La différence de moyenne estimée pour ces deux catégories ($\hat{\mu}_{\text{knee}} - \hat{\mu}_{\text{control}}$) vaut -0.027 , avec un intervalle de confiance à 95% pour cette différence qui vaut $[-0.95; 0.90]$. Cet intervalle, qui rassemble les valeurs les plus probables pour cette différence de moyenne, contient la valeur 0, ce qui confirme qu'il n'y a aucune raison de penser qu'une différence réelle existe entre ces 2 catégories. Le faible écart de moyennes observé entre ces 2 groupes est donc très vraisemblablement le fruit du hasard. L'éclairage du creux poplité donne les mêmes résultats que quand les patients sont maintenus dans le noir.

En revanche, les lignes 2 et 3 de ce tableau montrent des différences significatives ($p = 0.008$ et $p = 0.012$ pour les comparaisons **eyes/control** et **eyes/knee** respectivement). Les différences sont négatives, de l'ordre de -1.2 pour les 2 comparaisons, ce qui traduit des valeurs plus faibles pour **eyes** que pour les 2 autres groupes. Pour ces 2 comparaisons, les intervalles de confiance à 95% des différences ne contiennent pas le 0, mais exclusivement des valeurs négatives. Cela traduit donc bien une resynchronisation plus rapide chez les sujets dont les yeux sont exposés à la lumière que chez les sujets des 2 autres groupes.

En utilisant le tableau ci-dessus, nous pouvons synthétiser graphiquement ces résultats :

```
tidy(post_hoc2) %>%  
  ggplot(aes(x = contrast, y = estimate)) +
```

```
geom_point() +
geom_linerange(aes(ymin = conf.low, ymax = conf.high)) +
geom_hline(yintercept = 0, linetype = 2) +
labs(x = "Comparaison",
      y = "Différence de moyennes (et IC 95%)") +
coord_flip() +
theme_bw()
```



Notez ici l'utilisation de `geom_linerange`, pour afficher les intervalles de confiance à 95% des différences de moyennes. Il s'agit d'une alternative à `geom_errorbar()` dont nous avons déjà parlé [dans ce chapitre](#) du livre en ligne du semestre 4. La fonction `geom_hline()` permet de faire apparaître des lignes horizontales sur un graphique. Ici, avec `y = 0`, cette fonction fait apparaître un axe horizontal (axe des abscisses). Enfin, la fonction `coord_flip()` permet d'inverser les axes du graphique : l'axe des `x` bascule à la verticale, et l'axe des `y` à l'horizontale. Cela permet d'obtenir un graphique dont l'apparence est typique de ce genre de graphique produit avec les résultats du test de Tukey HSD.

Ce graphique montre bien que pour la comparaison **knee - control**, le zéro est compris dans l'intervalle de confiance à 95% de la différence de moyennes, ce qui confirme l'absence de différence significative de décalage de phase entre ces 2 groupes. À l'inverse, pour les 2 autres comparaisons (**eyes - knee** d'une part, et **eyes - control** d'autre part), les intervalles de confiance à 95% des différences de moyennes ne

coupent pas le zéro. Cela indique une différence de moyenne significative : dans la population générale, le zéro ne fait pas partie des valeurs les plus probables pour la différence de décalage de phase entre ces groupes.

Nous avons donc bien montré ici que la re-synchronisation de l'horloge interne n'est possible que par le biais de l'exposition des yeux à la lumière, et non du creux poplité.

1.7 L'alternative non paramétrique

1.7.1 La robustesse de l'ANOVA

Dans la suite de cette section, nous faisons l'hypothèse, bien que ça ne soit pas le cas, que les conditions d'application de l'ANOVA ne sont pas vérifiées pour notre jeu de données. Si les conditions d'application de l'ANOVA ne sont pas remplies, alors, les résultats de l'ANOVA ne peuvent pas être examinés car ils ne sont pas valides. Il nous faut alors recourir à un test non-paramétrique afin de comparer la moyenne de plus de deux groupes à la fois.

La particularité de l'ANOVA est sa grande **robustesse** vis-à-vis d'un non respect modéré de ses conditions d'application (voir définition de la robustesse dans la [?@sec-robust](#)). L'ANOVA étant particulièrement robuste, ses résultats resteront valides dans les situations suivantes :

- Non normalité modérée des résidus. Si les résidus ne suivent pas parfaitement une loi Normale mais qu'ils sont néanmoins grossièrement distribués selon une courbe en cloche, les résultats de l'ANOVA resteront vrais, surtout si les effectifs sont importants.
- Non homogénéité des résidus. Si les résidus ne sont pas homogènes dans tous les groupes, les résultats de l'ANOVA resteront vrais tant que les échantillons seront grands, approximativement de la même taille dans tous les groupes, et à condition que les écarts de variances entre les groupes ne dépassent pas un facteur 10.

Dans tous les autres cas de non respect des conditions de l'ANOVA, par exemple, si la variance des résidus n'est pas

homogène et que les groupes sont de petite taille ou de taille différente, ou si les variances diffèrent de plus d'un facteur 10, ou si les résidus s'écartent fortement de la normalité, ou **si les deux conditions d'application ne sont pas respectées (même modérément) en même temps**, il faudra alors faire un test non paramétrique.

L'alternative non paramétrique à l'ANOVA à un facteur est le test de la somme des rangs de **Kruskal-Wallis**

! Paramétrique ou non ?

Pour comparer la moyenne de plus de 2 groupes :

- lorsque les conditions permettant de réaliser un test paramétrique sont réunies (voir Section 1.6.2), on effectuera une **ANOVA**, qui n'est qu'une extension du test de **Student**. Si (et seulement si) on en rejette l'hypothèse nulle, on fera ensuite un test *post-hoc* **paramétrique** : le test de **comparaisons multiples de Student** et/ou le test de **Tukey HSD**.
- lorsque les conditions permettant de réaliser un test paramétrique ne sont pas réunies, on effectuera un test de **Kruskal-Wallis**, qui est une extension du test de **Wilcoxon**. Si (et seulement si) on en rejette l'hypothèse nulle, on fera ensuite un test *post-hoc* **non paramétrique** : le test de **comparaisons multiples de Wilcoxon** et/ou le test de **Dunn**.

1.7.2 Réalisation du tests et interprétation

Les hypothèses nulle et alternative du test de Kruskal-Wallis sont les suivantes. Comme toujours, l'hypothèse nulle concerne l'absence d'effet du facteur étudié :

- H_0 : le type de traitement appliqué n'a pas d'effet sur le décalage de phase. Les médianes sont égales dans tous les groupes ($med_{control} = med_{knee} = med_{eyes}$).
- H_1 : le type de traitement appliqué a un effet sur le décalage de phase. Les médianes ne sont pas toutes égales, au moins l'une d'entre elles diffère des autres.

La syntaxe du test est similaire à celle de l'ANOVA. On utilise la notation formule en plaçant la variable numérique expliquée à gauche du \sim , et le facteur (variable explicative) à droite du \sim :

```
kruskal.test(shift ~ treatment, data = Light)
```

Kruskal-Wallis rank sum test

data: shift by treatment

Kruskal-Wallis chi-squared = 9.4231, df = 2, p-value = 0.008991

Ici, la p -value est inférieure à α , on rejette donc H_0 : toutes les médianes ne sont pas égales. Comme avec l'ANOVA, il nous faut maintenant déterminer quelles médianes sont différentes les unes des autres, et quelles sont les magnitudes de ces différences. Pour cela, nous devons réaliser des tests *post-hoc* de comparaisons multiples.

1.7.3 Tests *a posteriori* ou tests *post-hoc*

Comme pour les tests *post-hoc* de l'ANOVA, nous allons voir ici 2 tests de comparaisons multiples non paramétriques.

1.7.3.1 Comparaisons multiples de Wilcoxon

Le premier test est l'équivalent non paramétrique du test de comparaisons multiples de Student : le **test de comparaisons multiples de la somme des rangs de Wilcoxon**. Le principe est absolument le même que pour le test de comparaisons multiples de Student : toutes les comparaisons 2 à 2 sont effectuées au moyen d'un test de la somme des rangs de Wilcoxon. Les p -values de ces tests sont corrigées afin de garantir que le risque d'erreur α global soit maintenu constant en dépit de l'augmentation du nombre de tests réalisés. Pour chaque comparaison, les hypothèses sont les suivantes :

- H_0 : la médiane des deux populations est égale.
- H_1 : la médiane des deux populations est différente.

```
# Réalisation du test
post_hoc3 <- pairwise.wilcox.test(Light$shift, Light$treatment)

# Affichage des résultats
post_hoc3
```

Pairwise comparisons using Wilcoxon rank sum exact test

data: Light\$shift and Light\$treatment

```
      control knee
knee 0.9551  -
eyes 0.0037  0.0524
```

P value adjustment method: holm

```
# Utilisation de la fonction `tidy` pour afficher les résultats dans un tibble
tidy(post_hoc3)
```

```
# A tibble: 3 x 3
  group1 group2 p.value
  <chr>   <chr>   <dbl>
1 knee   control 0.955
2 eyes   control 0.00373
3 eyes   knee    0.0524
```

Ici, la p -value du premier test est supérieure à $\alpha = 0.05$. Il n'y a donc pas de différence entre les médianes du groupe **control** et du groupe **knee**. Le traitement lumineux appliqué dans le creux poplité n'a donc aucun effet sur le décalage de phase.

La p -value du second test est en revanche inférieure à α . On rejette l'hypothèse nulle pour ce test ce qui confirme que le traitement lumineux appliqué au niveau des yeux a un effet sur le décalage de phase. Reste toutefois à quantifier l'importance de ce décalage de phase par rapport au groupe **control**.

Enfin, la p -value du troisième test est supérieure (tout juste !) à α . La conclusion logique est donc qu'il n'y a pas de différence significative entre les médianes des groupes **knee** et **eyes**. On sait que ce n'est pas le cas puisque nous avons montré plus haut (avec les tests paramétriques), que la différence de moyennes entre ces deux populations était significative. Nous avons ici l'illustration parfaite de la faible puissance des tests non paramétriques : leur capacité à détecter un effet lorsqu'il y en a réellement un est plus faible que celle des tests paramétriques. En outre, les procédures de comparaisons multiples sont très conservatives, et font mécaniquement baisser la puissance des tests pour maintenir constante l'erreur α . Je ne peux donc que vous inciter à la prudence lorsque vous interprétez les résultats d'un test de comparaisons multiples (*a fortiori* un test non paramétrique) pour lequel la p -value obtenue est très proche du seuil α .

Comme pour son homologue paramétrique, le test de comparaisons multiples de Wilcoxon nous permet de prendre une décision par rapport à H_0 , mais il ne nous dit rien de la magnitude des effets mesurés. Pour les connaître, il nous faut réaliser un autre test.

1.7.3.2 Le test de Dunn

Le **test de Dunn** est au test de Kruskal-Wallis ce que le test de Tukey HSD est à l'ANOVA : un test post-hoc permettant de déterminer la magnitude des effets observés. Pour pouvoir le réaliser, le package **DescTools** doit être chargé. Sa syntaxe est la même que pour le test de Kruskal-Wallis ou l'ANOVA :

```
# Réalisation du test
post_hoc4 <- DunnTest(shift ~ treatment, data = Light)

# Affichage des résultats
post_hoc4
```

```
Dunn's test of multiple comparisons using rank sums : holm
```

```
mean.rank.diff    pval
```

```
knee-control      -0.4821429 0.8859
eyes-control      -9.3392857 0.0164 *
eyes-knee         -8.8571429 0.0214 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec ces résultats on progresse un peu, car outre les p -values pour chaque comparaison, le test nous fournit une estimation de la différence des rangs moyens. Malheureusement, ces estimations sont souvent difficiles à interpréter (par exemple, quelle est l'unité utilisée ?) et aucun intervalle de confiance n'est fourni. On constate néanmoins que le test de Dunn donne ici des résultats comparables à ceux fournis par les tests paramétriques : le groupe **eyes** est significativement différent des deux autres. Pour obtenir les intervalles de confiance dont nous avons besoin, nous n'avons pas d'autre choix que des les calculer à l'aide du test de Wilcoxon classique, en réalisant manuellement les tests dont nous avons besoin. Ici, le test à proprement parler ne nous intéresse pas, d'ailleurs, sa p -value ne doit surtout pas être prise en compte car elle ignore totalement les comparaisons multiples et conduirait donc à augmenter l'erreur de type I. La seule chose pertinente est ici **la différence de (pseudo-)médiane estimée et son intervalle de confiance** :

```
# Comparaisons entre les groupes `knee` et `eyes` (dans cet ordre)
Light %>%
  filter(treatment %in% c("knee", "eyes")) %>%
  wilcox.test(shift ~ treatment, data = ., conf.int = TRUE) %>%
  tidy()

# A tibble: 1 x 7
  estimate statistic p.value conf.low conf.high method alternative
  <dbl>      <dbl>   <dbl>   <dbl>   <dbl> <chr>          <chr>
1     1.19         42 0.0262     0.3     2.21 Wilcoxon rank sum e~ two.sided
```

Pour le décalage de phase de ces 2 groupes, la différence de médiane estimée vaut donc 1.19, avec un intervalle de confiance à 95% de [0.3; 2.21]. Toutes les valeurs comprises dans cet intervalle de confiance sont strictement positives. Il y a donc très peu de chances pour que la différence de médiane entre ces deux groupes soit nulle. Le test de Dunn ci-dessus,

qui montre une différence significative entre ces groupes, est donc confirmé.

1.8 Exercices d'application

1.8.1 *Cardamine pensylvanica*

En biologie de la conservation, la question de l'existence d'un lien entre la capacité de dispersion des organismes et le maintien durable des populations dans le temps est étudié de près, notamment en raison de l'anthropisation des milieux qui conduit très souvent à la fragmentation des habitats. Cette question a été étudiée par 2 chercheurs (Molofsky et Ferdy 2005) chez *Cardamine pensylvanica*, une [plante annuelle d'Amérique du Nord](#) qui produit des graines qui sont dispersées de façon explosive. Quatre traitements ont été utilisés pour modifier expérimentalement la dispersion des graines. La distance entre populations contigües a été définie comme suit :

- Traitement 1 : **continu**. Les plants sont conservés au contact les uns des autres.
- Traitement 2 : **medium**. Les plants sont séparés de 23.2 centimètres.
- Traitement 3 : **long**. Les plants sont séparés de 49.5 centimètres.
- Traitement 4 : **isole**. Les plants sont séparés par des panneaux de bois empêchant la dispersion des graines.

Ces traitements ont été assignés au hasard à des populations de plantes, et 4 réplicats ont été faits pour chacun d'entre eux. Les résultats de l'expérience sont présentés ci-dessous. Il s'agit du nombre de générations durant lesquelles les plantes ont persisté :

- **continu** : 9, 13, 13, 16
- **medium** : 14, 12, 16, 16
- **long** : 13, 9, 10, 11
- **isole** : 13, 8, 8, 8

Saisissez ces données dans RStudio et faites-en l'analyse. Vous tenterez de déterminer si l'éloignement entre les populations de plantes a un impact sur leur capacité de survie.

Comme toujours, avant de vous lancer dans les tests, vous prendrez le temps de décrire les données avec des statistiques descriptives et des représentations graphiques.

1.8.2 Insecticides

L'efficacité de 6 insecticides nommés A, B, C, D, E et F a été testée sur 6 parcelles agricoles. Chaque insecticide de cette liste a été appliqué sur une parcelle agricole choisie au hasard. Deux semaines plus tard, 12 plants ont été collectés dans chaque parcelle agricole et le nombre d'insectes toujours vivants sur chacun d'entre eux a été compté. Les résultats sont présentés dans le fichier [Insectes.csv](#). Importez ces données dans **RStudio** et faites-en l'analyse. Tous les insecticides ont-ils la même efficacité ? Si la réponse est non, quels sont les insecticides les plus (ou les moins) efficaces.

1.8.3 La longueur des nageoires des manchots femelles

Avec le jeu de données `penguins` du package `palmerpenguins`, comparez la longueur des nageoires des femelles des 3 espèces de manchots. Les femelles des 3 espèces ont-elles toutes des nageoires de longueur différentes, et quelle est la magnitude de ces éventuelles différences ?

2 Corrélation

2.1 Pré-requis

Comme pour chaque nouveau chapitre, je vous conseille de travailler dans un nouveau script que vous placerez dans votre répertoire de travail, et dans une nouvelle session de travail (Menu **Session** > **Restart R**). Inutile en revanche de créer un nouveau **Rproject** : vos pouvez tout à fait avoir plusieurs script dans le même répertoire de travail et pour un même **Rproject**. Comme toujours, consultez [le livre en ligne du semestre 3](#) si vous ne savez plus comment faire.

Si vous êtes dans une nouvelle session de travail (ou que vous avez quitté puis relancé **RStudio**), vous devrez penser à recharger en mémoire les packages utiles. Dans ce chapitre, vous aurez besoin d'utiliser :

- le **tidyverse** (Wickham 2023), qui comprend notamment le package **readr** (Wickham, Hester, et Bryan 2023), pour importer facilement des fichiers **.csv** au format **tibble**, le package **dplyr** (Wickham, François, et al. 2023), pour manipuler des tableaux, et le package **ggplot2** (Wickham, Chang, et al. 2023) pour les représentations graphiques.
- **skimr** (Waring et al. 2022), qui permet de calculer des résumés de données très informatifs.

```
library(tidyverse)
library(skimr)
```

Vous aurez également besoin des jeux de données suivants, qu'il vous faut donc télécharger dans votre répertoire de travail :

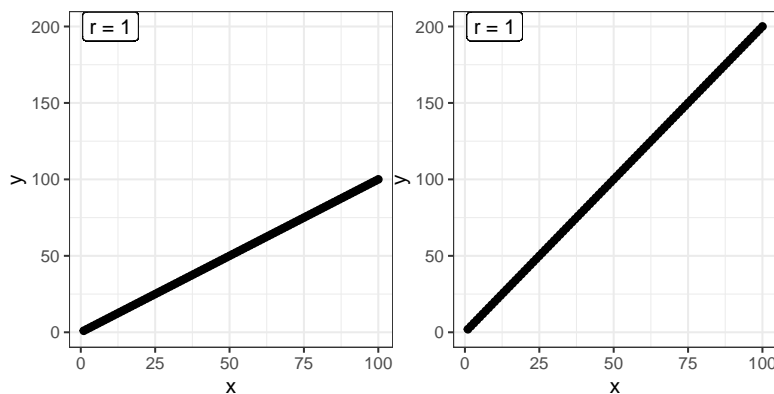
- [birds.csv](#)
- [loups.csv](#)
- [ropetrick.csv](#)

```
theme_set(theme_bw())
```

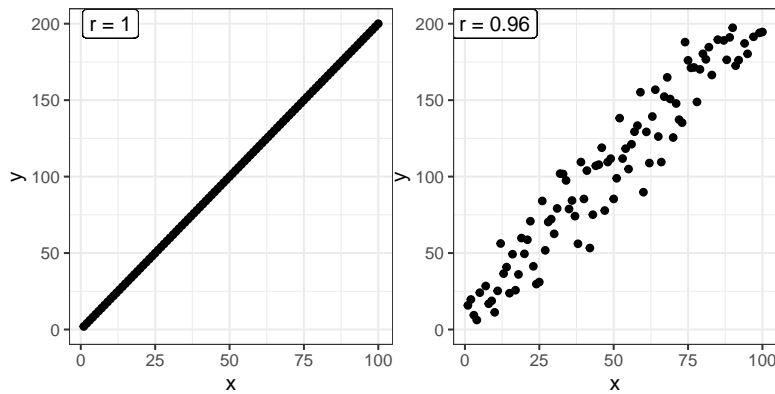
2.2 Principe

Lorsque des variables numériques sont associées on dit qu'elles sont **corrélées**. Par exemple, la taille du cerveau et la taille du corps sont corrélées positivement parmi les espèces de mammifères. Les espèces de grande taille ont tendance à avoir un cerveau plus grand et les petites espèces ont tendance à avoir un cerveau plus petit. Le coefficient de corrélation est la quantité qui décrit la force et la direction de l'association entre deux variables numériques mesurées sur un échantillon de sujets ou d'unités d'observation. La corrélation reflète la quantité de dispersion dans un nuage de points entre deux variables. Contrairement à la régression linéaire, la corrélation n'ajuste aucune droite à des données et ne permet donc pas de mesurer à quel point le changement d'une variable entraîne un changement rapide ou lent de l'autre variable.

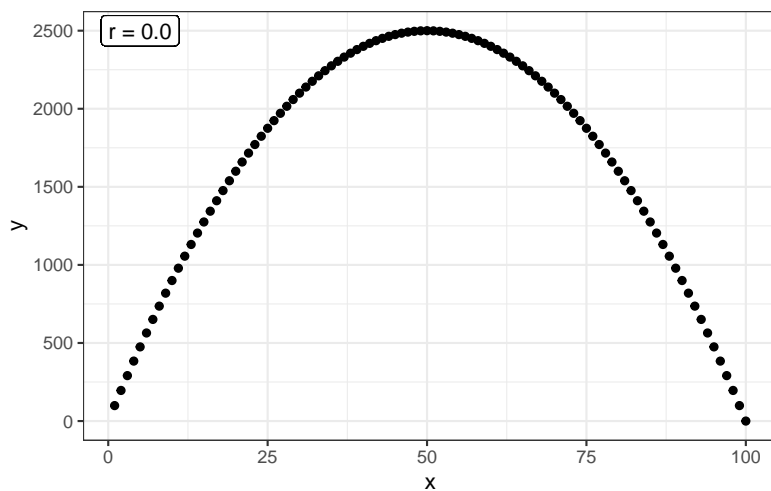
Ainsi, sur la figure ci-dessous, le coefficient de corrélation entre X et Y est le même pour les deux graphiques : il vaut 1.



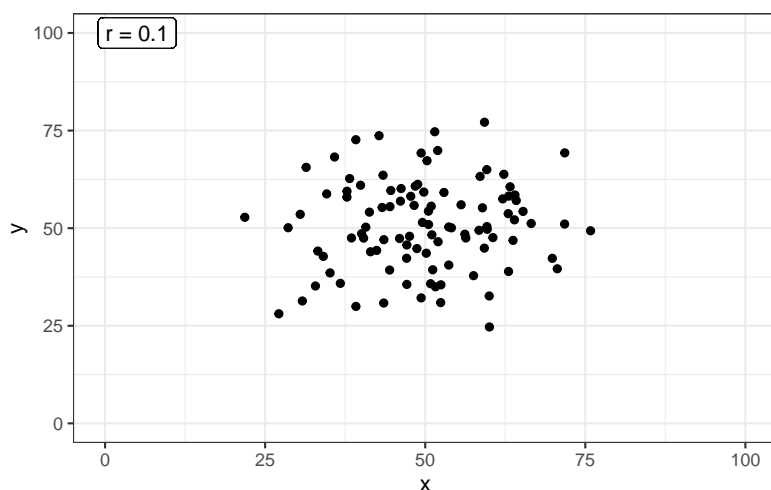
Ici, le coefficient de corrélation (noté r) vaut 1 dans les deux cas, car tous les points sont alignés sur une droite. La pente de la droite n'influence en rien la valeur de corrélation. En revanche, le degré de dispersion des points autour d'une droite parfaite a une influence :



Plus la dispersion autour d'une droite parfaite sera grande, plus la corrélation sera faible. C'est la raison pour laquelle lorsque l'on parle de "corrélation", on sous-entend généralement **corrélation linéaire**. Ainsi, 2 variables peuvent avoir une relation très forte, mais un coefficient de corrélation nul, si leur relation n'est pas linéaire :



L'exploration graphique de vos données devrait donc toujours être une priorité. Calculer un coefficient de corrélation nul ou très faible ne signifie par pour autant une absence de relation entre les 2 variables numériques étudiées. Cela peut signifier une relation non linéaire. La solution la plus simple pour distinguer une relation telle que celle du graphique précédent, et une absence de relation telle que celle présentée dans le graphique ci-dessous, est l'examen visuel des données :



En bref, le coefficient de corrélation r est compris entre -1 et +1 :

- Une forte valeur absolue (r proche de -1 ou +1), indique une relation presque linéaire.
- Une faible valeur absolue indique soit une absence de relation, soit une relation non linéaire (la visualisation graphique permet généralement d'en savoir plus).
- Une valeur positive indique qu'une augmentation de la première variable est associée à une augmentation de la seconde variable.
- Une valeur négative indique qu'une augmentation de la première variable est associée à une diminution de la seconde variable.

! Important

Le coefficient de corrélation suppose une relation linéaire entre les deux variables numériques examinées. Calculer un coefficient de corrélation très faible peut indiquer :

- une absence de relation entre les variables étudiées
- une relation forte, mais non linéaire entre les variables étudiées

La façon la plus simple de distinguer ces 2 cas de figure très différents est l'**exploration graphique des données**.

Dans la suite de ce chapitre, nous allons voir comment calculer le coefficient de corrélation entre 2 variables numériques¹, et puisque nous travaillons avec des **échantillons**, ce calcul sera nécessairement entaché d'**incertitude**. Tout comme la moyenne ou la variance d'un échantillon, la corrélation est un **paramètre** des populations dont nous ne pourrions qu'estimer la valeur. Toute estimation de corrélation devra donc être encadrée par un intervalle d'incertitude, généralement, il s'agit de l'intervalle de confiance à 95% de la corrélation. Enfin, outre l'estimation de la valeur de la corrélation et de son incertitude, nous pourrions aussi faire des tests d'hypothèses au sujet des corrélations que nous estimerons. En particulier, nous pourrions tester si la corrélation observée est significativement différente de zéro ou non.

¹ Vous aurez compris je pense qu'un calcul de corrélation n'a de sens que si l'on dispose de 2 variables numériques, enregistrées sur les mêmes individus ou unités d'étude. Voir détails à la fin de la Section 2.4

2.3 Contexte

Les adultes qui infligent des mauvais traitements à leurs enfants ont souvent été maltraités dans leur enfance. Une telle relation existe-t-elle également chez d'autres espèces animales, chez qui cette relation pourrait être étudiée plus facilement ? Müller et al. (2011) ont étudié cette possibilité chez le **fou de Grant** (*Sula granti*), un oiseau marin colonial vivant entre autres aux Galápagos. Les jeunes laissés au nid sans attention parentale reçoivent fréquemment la visite d'autres oiseaux, qui se comportent souvent de manière agressive à leur rencontre. Les chercheurs ont compté le nombre de ces visites dans le nid de 24 poussins dotés d'une bague d'identification individuelle. Ces 24 individus ont ensuite été suivis à l'âge adulte, lorsqu'ils sont à leur tour devenus parents. On cherche donc à savoir s'il existe un lien entre le nombre de visites agressives qu'un individu a reçu lorsqu'il était à l'état de poussin, et un degré d'agressivité mesuré à l'âge adulte.

2.4 Importation et mise en forme des données

Les données récoltées par les chercheurs figurent dans le fichier `birds.csv`. Importez ces données dans RStudio dans

un objet noté `birds`.

```
birds

# A tibble: 24 x 2
  nVisitsNestling futureBehavior
      <dbl>         <dbl>
1             1         -0.8
2             7        -0.92
3            15         -0.8
4             4        -0.46
5            11        -0.47
6            14        -0.46
7            23        -0.23
8            14        -0.16
9             9        -0.23
10           5        -0.23
# i 14 more rows
```

La première colonne de ce tableau indique, pour chaque individu suivi, le nombre de visites reçues au nid de la part d’adultes agressifs lorsqu’ils étaient poussins. La seconde colonne indique, pour ces mêmes individus devenus adultes, le nombre de visites agressives effectuées à des nids d’autres poussins. Ce nombre n’est pas dans la même unité que la première variable car il a été corrigé par d’autres variables d’intérêt pour les chercheurs.

Il manque à ce tableau une variable indiquant le code des individus. Elle n’est pas indispensable, mais la rajouter est une bonne habitude à prendre pour toujours travailler avec des “données rangées”. Puisqu’on dispose de 24 individus, on leur assigne donc un code de 1 à 24 :

```
birds <- birds %>%
  mutate(ID = factor(1:24))
birds

# A tibble: 24 x 3
  nVisitsNestling futureBehavior ID
      <dbl>         <dbl> <fct>
1             1         -0.8  1
```

2	7	-0.92	2
3	15	-0.8	3
4	4	-0.46	4
5	11	-0.47	5
6	14	-0.46	6
7	23	-0.23	7
8	14	-0.16	8
9	9	-0.23	9
10	5	-0.23	10

i 14 more rows

Présentées sous cette forme, les données ressemblent beaucoup à celles du **?@sec-moy2**. Ça n'est pas un hasard : les données dont nous disposons ici sont appariées. Calculer la corrélation entre 2 variables n'a de sens que si chaque unité d'échantillonnage ou d'observation (ici, les individus), fournissent 2 valeurs dont on souhaite mesurer l'association. Dans l'étude sur les effets de la testostérone chez les carouges à épaulettes, on avait, pour chaque individu étudié, 2 mesures d'immunocompétence : une avant et l'autre après l'opération chirurgicale. Ici, chaque Fou de Grant étudié fournit 2 valeurs également. Contrairement à l'étude des carouges à épaulettes, il s'agit de deux variables distinctes (nombre de visites agressives reçues à l'état de poussin d'une part, et comportement agressif à l'âge adulte d'autre part), mais les 2 mesures sont bien liées puisqu'elles sont obtenues chez le même individu.

Pour bien enfoncer le clou, voici un autre exemple. Calculer la corrélation entre la taille des femmes françaises et la tension artérielle des femmes anglaises n'a strictement aucun sens car ce sont des groupes de femmes distincts qui fournissent les mesures de chaque variable. En revanche, sélectionner un groupe de femmes au hasard dans la population mondiale, et examiner, pour chacune des femmes de l'échantillon, à la fois la taille et la tension artérielle est pertinent. On peut alors se poser la question de lien potentiel existant entre ces 2 variables dans la population générale. L'étude de la corrélation entre la taille et la tension artérielle chez les femmes prend alors tout son sens.

! Important

Calculer une corrélation n'a de sens que si les données étudiées sont appariées.

2.5 Exploration statistique des données

Comme toujours, la première chose à faire est d'examiner quelques statistiques descriptives pour se faire une idée de la forme des données et pour repérer les éventuelles données manquantes ou aberrantes.

```
skim(birds)
```

```
-- Data Summary -----
```

	Values
Name	birds
Number of rows	24
Number of columns	3

```
-----  
Column type frequency:
```

factor	1
numeric	2

```
-----  
Group variables      None
```

```
-- Variable type: factor -----
```

	skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
1	ID	0	1	FALSE	24	1: 1, 2: 1, 3: 1, 4: 1

```
-- Variable type: numeric -----
```

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
1	nVisitsNestling	0	1	13.1	7.21	1	8.75	13	15.8
2	futureBehavior	0	1	-0.119	0.374	-0.92	-0.288	-0.1	0.182

p100 hist

1	31
2	0.39

Outre le facteur ID que nous venons de créer, nous disposons donc de 2 variables numériques qui ne contiennent pas de

données manquantes.

1. La variable `nVisitsNestling`, qui indique le nombre de visites agressives reçues par les individus suivis lorsqu'ils étaient de jeunes poussins, varie de 1 à 31, pour une moyenne de 13.12, une médiane proche (13) mais un écart-type important.
2. La variable `futureBehavior` varie de -0.92 à 0.39, avec une moyenne et une médiane proche de 0 (-0.12 et -0.1 respectivement).

Comme toujours, la fonction `skim()` nous renseigne sur la tendance centrale, ou **position**, des variables étudiées (grâce aux moyennes et médianes) et sur la **dispersion** des données (grâce à l'écart-type et aux "minis-histogrammes"). Ici, si la variable `nVisitsNestling` semble être à peu près distribuée selon une courbe en cloche (asymétrique), ce n'est pas le cas de la variable `futureBehavior` qui semble présenter une très forte asymétrie à gauche.

D'habitude, on calcule à ce stade des indices d'**incertitude** : l'erreur standard de la moyenne ou l'intervalle de confiance de la moyenne. Ici, ça n'est pas utile car les moyennes en elles-mêmes ne nous intéressent pas, et donc leurs incertitudes non plus. C'est en revanche la relation entre les 2 variables numériques qui nous intéresse, en particulier l'intensité et le sens de cette relation. On calcule donc maintenant le coefficient de corrélation linéaire entre les 2 variables :

```
birds %>%  
  select(nVisitsNestling, futureBehavior) %>%  
  cor()
```

	nVisitsNestling	futureBehavior
nVisitsNestling	1.0000000	0.5337225
futureBehavior	0.5337225	1.0000000

Le résultat est fourni sous la forme d'une matrice symétrique :

- Sur la diagonale, les corrélations valent 1 (le coefficient de corrélation d'une variable avec elle-même vaut toujours 1).

- En dehors de la diagonale, on trouve le coefficient de corrélation linéaire entre les 2 variables d'intérêt. Ici, il est positif et vaut 0.534, ce qui est une valeur relativement élevée dans le domaine de la biologie ou de l'écologie. Le signe positif de la corrélation indique que lorsque la première variable augmente, la seconde variable augmente également. Autrement dit, plus les fous de Grant ont été maltraités quand ils étaient poussins, plus ils adoptent un comportement agressif à l'âge adulte.

2.6 Exploration graphique des données

Pour répondre à la question posée et visualiser la relation entre les deux variables numériques, on peut simplement associer chaque variable à un axe d'un graphique et faire un nuage de points. Je vous encourage à jeter un œil à [ce chapitre du livre en ligne du semestre 3](#) pour voir quels types de graphiques sont pertinents dans cette situation.

Afin de savoir si la valeur de r calculée précédemment dans notre échantillon (0.534) reflète une relation linéaire mais moyenne, ou une relation qui n'est pas vraiment linéaire, nous pouvons donc faire un nuage de points :

```
birds %>%  
  ggplot(aes(x = nVisitsNestling, y = futureBehavior)) +  
  geom_point() +  
  labs(x = "Nombre de visites reçues par le poussin",  
       y = "Agressivité à l'âge adulte")
```

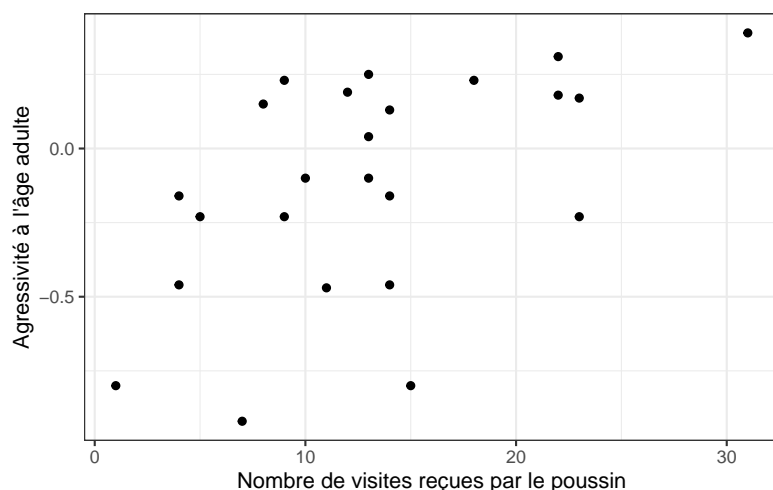


Figure 2.1: Relation entre agressivité à l'âge adulte et nombre de visites agressives reçues par les poussins de l'espèce *Sula granti*

On constate ici que la corrélation moyenne obtenue plus haut est due au fait que les points sont assez dispersés, et non au fait que la relation n'est pas linéaire. On peut donc dire que la relation, si elle existe, n'est pas parfaite. Le comportement des individus devenus adultes semble donc en partie lié au nombre de visites agressives qu'ils ont reçues étant jeunes, mais ce n'est certainement pas le seul facteur influençant leur comportement. Un test d'hypothèses devrait nous permettre de déterminer si la corrélation linéaire observée ici est simplement le fruit du hasard de l'échantillonnage, ou si au contraire la relation observée n'est pas seulement le fruit du hasard, mais bien le reflet d'un lien réel entre les 2 variables.

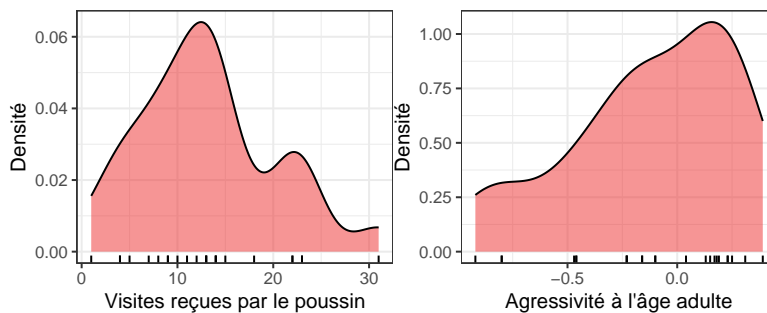
Si visualiser la distribution des données n'est pas indispensable pour se faire une idée de la nature du lien qui existe (ou non) entre les deux variables, cela sera néanmoins utile pour vérifier les conditions d'application du test de corrélations paramétrique de Pearson. Comme dans les chapitres précédents, nous avons donc intérêt à examiner la distribution de ces 2 variables par le biais d'histogrammes, de graphiques de densité ou de boîtes à moustaches.

```
birds %>%
  ggplot(aes(x = nVisitsNestling)) +
  geom_density(fill = "firebrick2", alpha = 0.5) +
```

```

geom_rug() +
labs(x = "Visites reçues par le poussin",
     y = "Densité")
birds %>%
ggplot(aes(x = futureBehavior)) +
geom_density(fill = "firebrick2", alpha = 0.5) +
geom_rug() +
labs(x = "Agressivité à l'âge adulte",
     y = "Densité")

```



Aucune des 2 variables ne semble suivre parfaitement une distribution Normale. Il faudra réaliser des tests de normalité pour en avoir le cœur net.

2.7 Le test paramétrique

2.7.1 Les hypothèses

Comme pour la plupart des grandeurs calculées à partir d'un échantillon, la corrélation r n'est qu'un estimateur de la corrélation qui existe réellement entre ces deux variables dans la population générale. Dans la population générale, la corrélation linéaire est généralement notée ρ . Son estimateur, r est donc souvent noté $\hat{\rho}$.

Le test d'hypothèses que nous allons faire maintenant permet de vérifier si le coefficient de corrélation ρ dans la population générale est différent de 0 ou non. Les hypothèses de ce test sont les suivantes :

- H_0 : le coefficient de corrélation entre les deux variables étudiées vaut 0 dans la population générale ($\rho = 0$). Autrement dit, la corrélation observée dans l'échantillon

n'est que le fruit du hasard de l'échantillonnage : il n'y a aucun lien entre les 2 variables dans la population générale.

- H_1 : le coefficient de corrélation entre les deux variables étudiées est différent de 0 dans la population générale ($\rho \neq 0$). La fluctuation d'échantillonnage ne suffit pas à expliquer la corrélation observée : en plus du hasard de l'échantillonnage, il existe bel et bien un lien entre les 2 variables étudiées.

Ce test est réalisé dans **RStudio** grâce à la fonction `cor.test()`, qui permet, selon les arguments renseignés, de réaliser soit :

- le test de corrélation paramétrique de Pearson.
- le test de corrélation non paramétrique de Spearman.

2.7.2 Conditions d'application

Comme toujours, on cherche à réaliser un test paramétrique (ici, le test de Pearson) si les données le permettent. Pour avoir le droit de réaliser le test de corrélation de Pearson, il nous faut donc en vérifier les conditions d'application :

1. Les individus doivent être indépendants les uns des autres
2. Les mesures effectuées doivent suivre une **distribution Normale bivariée**

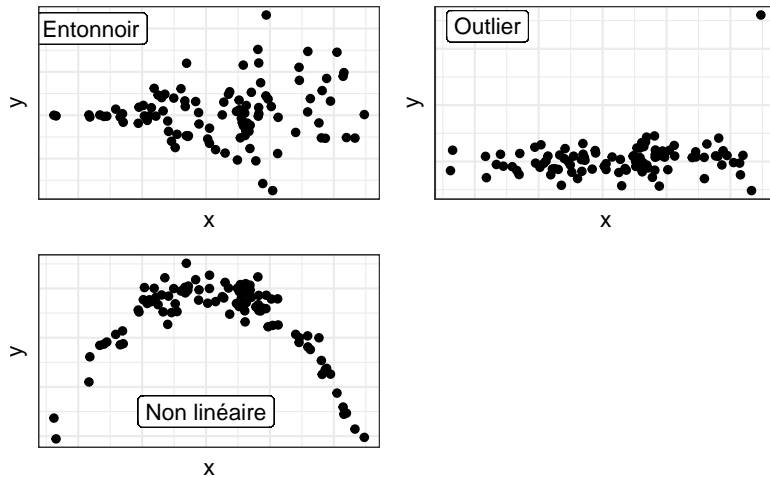
Comme toujours, sauf si on a de bonnes raisons de penser le contraire, on considère généralement que si l'échantillonnage a été fait de façon aléatoire, l'indépendance des observations est garantie. La condition de "distribution Normale bivariée" des données est en revanche nouvelle. Elle suppose essentiellement que les 3 critères suivants soient vérifiés :

1. La relation entre les 2 variables doit être linéaire. C'est que nous tentons de vérifier visuellement en réalisant un nuage de points des données.
2. Sur un graphique représentant une variable en fonction de l'autre, le nuage de points doit avoir une forme circulaire ou elliptique. Là encore, une représentation graphique nous permet d'apprécier cette condition.

3. Les 2 variables étudiées doivent suivre une distribution Normale dans la population générale. Avant de faire ce test, il nous faut donc vérifier la Normalité des données pour chacune des 2 variables séparément, à l'aide, par exemple, d'un test de Shapiro-Wilk.

Pour résumer, l'examen du nuage de points permet de vérifier les 2 premières conditions et 2 tests de Shapiro permettent de vérifier la troisième. Pour l'examen du nuage de points, les conditions ne seront pas remplies dans les situations suivantes (voir les exemples du graphique ci-dessous) :

- Le nuage de points a une forme d'entonnoir ou de nœud papillon.
- Des outliers sont présents (quelques points fortement éloignés du reste des observations).
- Une relation non linéaire existe entre les deux variables.



Enfin, si l'une, l'autre ou les deux séries de données ne suivent pas la loi Normale, il faudra faire un test non paramétrique.

Dans notre cas, le graphique Figure 2.1 semble indiquer que les 2 premières conditions d'application sont remplies (la relation entre les deux variable semble globalement linéaire et le nuage de points a globalement une forme elliptique). Il nous reste donc à vérifier la normalité des 2 variables. Les hypothèses nulles et alternatives du test de Shapiro-Wilk sont toujours les mêmes :

- H_0 : les données suivent une distribution Normale dans la population générale.

- H_1 : les données ne suivent pas une distribution Normale dans la population générale.

```
birds %>%
  pull(nVisitsNestling) %>%
  shapiro.test()
```

Shapiro-Wilk normality test

```
data: .
W = 0.95783, p-value = 0.3965
```

Contrairement à ce que pouvait laisser croire le graphique de densité, la variable `nVisitsNestling` suit bien une distribution Normale (test de Shapiro-Wilk, $p = 0.397$).

```
birds %>%
  pull(futureBehavior) %>%
  shapiro.test()
```

Shapiro-Wilk normality test

```
data: .
W = 0.91575, p-value = 0.04709
```

En revanche, au seuil $\alpha = 0.05$, la variable `futureBehavior` ne suit pas une distribution Normale (test de Shapiro-Wilk, $p = 0.047$).

Les conditions d'application ne sont pas vérifiées. En toute rigueur, il nous faudrait donc réaliser ici le test non-paramétrique de Spearman. Nous verrons comment le faire plus tard. Pour l'instant, et pour que vous sachiez comment faire, nous allons faire comme si les conditions d'application du tests paramétrique étaient bel et bien remplies, et nous allons donc réaliser le test paramétrique de Pearson.

2.7.3 Réalisation du test et interprétation

La syntaxe du test est très simple :

```
cor.test(birds$nVisitsNestling, birds$futureBehavior)
```

Pearson's product-moment correlation

```
data: birds$nVisitsNestling and birds$futureBehavior
t = 2.9603, df = 22, p-value = 0.007229
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1660840 0.7710999
sample estimates:
      cor
0.5337225
```

Comme expliqué plus haut (et sur la première ligne des résultats du test), il s'agit du **test paramétrique de corrélation de Pearson**. Comme pour tous les tests examinés jusqu'ici, les premières lignes des résultats fournissent toutes les informations utiles au sujet du test. Ici, on peut dire :

Au seuil $\alpha = 0.05$, le test de corrélation de Pearson a permis de rejeter l'hypothèse nulle selon laquelle le nombre de visites agressives au nid des poussins et leur futur comportement agressif sont indépendants ($t = 2.96$, $ddl = 22$, $p = 0.007$).

Ce test prouve donc que ρ est significativement différent de 0. La valeur de 0.53 observée ici n'est pas due au seul hasard de l'échantillonnage.

Comme toujours, les résultats du test que nous avons réalisé ne nous disent rien de la valeur de la corrélation estimée, ni de son incertitude. Il nous faut pour cela examiner les autres lignes fournies par RStudio lorsque nous faisons ce test et qui relèvent de l'estimation (voir section suivante).

Dernière chose concernant ce test, nous avons fait ici un test bilatéral comme nous le rappelle cette ligne des résultats :

alternative hypothesis: true correlation is not equal to 0

Comme pour les tests de comparaisons de moyennes, il est possible de réaliser un test unilatéral, à condition que cela ait un sens, à condition que nous soyons en mesure d'expliquer le choix de notre hypothèse alternative. La syntaxe est la même que pour les tests de Student ou de Wilcoxon : on utilise l'argument `alternative = "less"` ou `alternative = "greater"` au moment de faire le test, selon l'hypothèse que l'on souhaite tester.

Ici, si les hypothèses que nous souhaitons tester sont les suivantes :

- H_0 : le coefficient de corrélation entre les deux variables étudiées vaut 0 dans la population générale ($\rho = 0$)
- H_1 : le coefficient de corrélation entre les deux variables étudiées est positif dans la population générale ($\rho > 0$)

On utilise la syntaxe suivante :

```
cor.test(birds$nVisitsNestling, birds$futureBehavior,  
         alternative = "greater")
```

Pearson's product-moment correlation

```
data: birds$nVisitsNestling and birds$futureBehavior  
t = 2.9603, df = 22, p-value = 0.003615  
alternative hypothesis: true correlation is greater than 0  
95 percent confidence interval:  
 0.2320921 1.0000000  
sample estimates:  
      cor  
0.5337225
```

Comme pour les autres test unilatéraux, le choix d'une hypothèse alternative aberrante se traduit par une p -value très forte, généralement égale à (ou très proche de) 1. Dans le cas précis de cette étude, il serait abusif de faire un tel test unilatéral. En effet, les scientifiques suppose que si un lien existe entre les deux variables, la corrélation devrait être positive.

Mais avoir observé une relation de cette nature chez d'autres espèces (qui plus est, chez des espèces très différentes) n'est pas suffisant. Car peut-être qu'une relation inverse peut également être observée dans d'autres groupes, et on peut très bien imaginer des mécanismes permettant de l'expliquer. Enfin, avoir observé une corrélation positive dans notre échantillon lors de l'examen préliminaire des données n'est jamais une raison suffisante pour choisir une hypothèse alternative unilatérale. Le choix des hypothèses devrait en effet toujours être effectué avant la collecte des données (voir ?@sec-bilat). Il est donc bien plus honnête de réaliser un test unilatéral, puis, en cas de rejet de H_0 , de revenir aux estimation pour interpréter les résultats et conclure. C'est que nous allons voir maintenant.

2.7.4 Estimation et intervalle de confiance

Revenons à notre test bilatéral. La section "estimation" des résultats de ce test nous indique que la meilleure estimation du coefficient de corrélation linéaire de Pearson dans la population générale vaut $\hat{\rho} = 0.533$. C'est la valeur que nous avons calculé à la main avec la fonction `cor()`.

L'intervalle de confiance à 95% de cette valeur estimée est également fourni. La conclusion de cette procédure pourrait donc être formulée de la façon suivante :

Au seuil $\alpha = 0.05$, le test de corrélation de Pearson a permis de rejeter l'hypothèse nulle selon laquelle le nombre de visites agressives au nid des poussins et leur futur comportement agressif sont indépendants ($t = 2.96$, $ddl = 22$, $p = 0.007$). La meilleure estimation du coefficient de corrélation dans la population générale vaut $\hat{\rho} = 0.533$. La vraie valeur dans la population générale a de bonnes chances de se trouver dans l'intervalle $[0.17 ; 0.77]$ (intervalle de confiance à 95%).

Autrement dit, le test a permis de rejeter l'hypothèse nulle et d'affirmer que les 2 variables sont corrélées. L'estimation du coefficient de corrélation et de son intervalle de confiance nous permettent de préciser le sens de cette relation (positive ou négative) et quantifier l'intensité de cette relation. Ici, la relation est bien positive : plus un individu est exposé à

des comportements agressif au stade de poussin, plus il aura tendance à reproduire de tels comportements à l'âge adulte. L'incertitude associée à cette estimation de coefficient de corrélation est très grande (IC95% : [0.17 ; 0.77]). Un échantillonnage plus large permettrait de le réduire. Mais les études de ce type sont très coûteuses, notamment en temps, et on est souvent obligé de se contenter des données dont on dispose. La vraie corrélation entre ces 2 variables pourrait donc être relativement faible dans la population générale (0.17), laissant supposer que l'agressivité à l'âge adulte est finalement peu liée à l'agressivité à laquelle les poussins ont été exposés. Mais elle pourrait aussi être très forte (0.77), laissant supposer que l'agressivité à l'âge adulte est fortement liée à l'exposition des poussins à des comportement agressif. On voit bien ici que le seul test de corrélation ne permet pas de trancher dans l'absolu. Tout ce que fait un test, c'est dire si oui ou non on dispose d'assez de preuve pour affirmer qu'une hypothèse nulle est fausse. Le reste de l'interprétation dépend de l'estimation des paramètres de la population générale et de leur incertitude. Quand l'incertitude est faible, on peut être assez affirmatif. Mais quand elle est forte, comme ici, il faut rester prudent quant aux interprétations possibles.

2.8 Corrélation et causalité

2.8.1 Quelques exemples évidents

On entend souvent que “**Corrélation n'est pas causalité**”. Cela signifie que la corrélation ne mesure qu'un lien entre 2 variables, mais pas nécessairement que les variations de la première influencent celles de la deuxième. En réalité, si on dispose d'un nombre de variables suffisamment grand, on pourra toujours en trouver 2 qui sont fortement corrélées, sans qu'il n'y ait la moindre relation de causalité entre les deux.

Par exemple, le nombre de morts par noyade aux états unis est corrélé à 66% ($r = 0.66$) avec le nombre de films dans lesquels Nicolas Cage joue chaque année. L'un n'est certainement pas la cause de l'autre. De même, le nombre de doctorats accordés chaque année dans le domaine du génie civil est

corrélé à 96% ($r = 0.959$) avec la consommation de mozzarella. Là encore, on ne voit pas bien quelle relation de cause à effet pourrait exister entre ces 2 variables. Ces exemples (et de nombreux autres) peuvent être retrouvés, chiffres à l'appui, [sur ce site web](#).

Pour ces exemples extrêmes, il est évident que la corrélation ne doit pas être interprétée comme une relation de cause à effet. Deux variables fortement corrélées sont simplement deux variables qui varient conjointement, dans le même sens (si la corrélation est positive) ou dans le sens opposé (si la corrélation est négative).

2.8.2 Les variables confondantes

Dans certaines situations, il est pourtant tentant de parler de causalité. Par exemple, à la fin des années 1990, des scientifiques ont montré, dans une étude tout à fait sérieuse, que dans les villes de France où l'on consomme le plus de crème solaire, la prévalence des cancers de la peau est également la plus forte. Certains journaux de vulgarisation scientifique se sont empressés de reprendre ce résultat (une corrélation positive entre utilisation de crème solaire et prévalence des mélanomes), et de conclure, à tort, que la crème solaire contribuait donc à donner le cancer de la peau. Pourtant, “corrélation n'est pas causalité” ! Une variable importante, pourtant évoquée dans l'article scientifique, est restée ignorée des journalistes scientifiques de l'époque : l'exposition au soleil. En effet, dans les villes où l'exposition au soleil est la plus forte (les villes de la côte méditerranéenne par exemple), on met en moyenne plus de crème solaire qu'ailleurs, mais on développe aussi plus de mélanomes qu'ailleurs. À l'inverse, dans les villes les moins ensoleillées de France, on utilise beaucoup moins de crème solaire, mais on développe aussi beaucoup moins de mélanomes, simplement parce qu'on est moins exposé au risque.

Dans ce dernier exemple, la variable **exposition au soleil** est une **variable confondante** (ou “confounding variable” en anglais). C'est elle qui cause à la fois l'augmentation de la prévalence des mélanomes, et l'augmentation de l'utilisation de crème solaire. On a donc bien 2 relations de causalité, mais pas entre les variables que l'on étudie. La corrélation que l'on observe entre **prévalence des mélanomes** et **utilisation**

de crème solaire n'est que la conséquence des relations de causalité avec la variable confondante.

La difficulté est ici que l'on ne peut pas savoir à l'avance quelle variable confondante pourrait venir influencer les variables que nous mesurons. Dans le cas de l'agressivité du fou de Grant, des traits génétiques particuliers pourraient par exemple expliquer le lien que nous observons entre nos 2 variables. Imaginons par exemple que la présence de certains allèles dans le génome des individus soit responsables à la fois d'une plus grande agressivité à l'âge adulte, et d'une plus grande autonomie lorsqu'ils sont jeunes. Des poussins possédant ces allèles seront plus autonomes que d'autres, il seront donc laissés plus souvent seuls par leurs parents, ce qui les exposera à des visites plus fréquentes d'adultes agressifs. Sous cette hypothèse, les 2 variables que nous avons étudiées ne sont liées entre elles que parce qu'il existe une relation de causalité entre les traits génétique des individus et chacune des 2 variables étudiées. C'est la raison pour laquelle, lorsque j'ai décrit les résultats de nos tests et des analyses descriptives, j'ai bien fait attention à ne pas dire que les visites agressives auprès des poussins étaient la cause de l'agressivité future des adultes. Je me suis contenté de dire que les variables étaient liées, et que plus un poussin reçoit de visites agressives, plus il aura lui même un comportement agressif à l'âge adulte. Prouver que l'un est la cause de l'autre, ou que l'autre est la conséquence de l'un est impossible avec ce type d'étude.

2.8.3 Études expérimentales ou observationnelles

Les exemples que nous venons d'aborder concernent tous des études dites **observationnelles**. À l'inverse des études **expérimentales**, dans lesquels l'expérimentateur a un certain contrôle des variables confondantes potentielles, ça n'est presque jamais le cas des études observationnelles. Dans l'exemple des fous de Grant, les scientifiques n'ont fait qu'observer des comportements dans une population naturelle. Ils n'ont pas eu la possibilité de vérifier en amont que tous les individus suivis avaient des gènes "normaux" vis-à-vis de l'agressivité. Dans le cas de l'étude sur la crème solaire, les chercheurs n'ont fait qu'observer ce qui se passe à plusieurs endroits de France. Ils n'ont pas pu s'assurer

que l'ensoleillement était le même dans toutes les villes sur laquelle a porté cette étude.

Les études expérimentales sont les seules à permettre d'établir des relations de cause à effet. C'est comme cela que par exemple, on peut affirmer qu'un vaccin est efficace ou non contre tel ou tel virus, ou à l'inverse qu'il présente tel ou tel effet secondaire. Pour tester l'efficacité d'un vaccin vis-à-vis d'un virus spécifique, on met en place une étude expérimentale dite "en double aveugle". Dans la population générale, on va constituer 2 échantillons de patients atteints par le virus. On administrera ensuite le vaccin à l'un des deux groupes, alors qu'on distribuera un placebo (le même vaccin mais sans son composé actif) à l'autre groupe, dans les mêmes conditions. L'étude est "en double aveugle", car les patients ne savent pas s'ils reçoivent le vaccin actif ou inactif, et les médecins qui administrent le traitement non plus. Cela a pour but de contrôler l'effet placebo.

Dans ces études, la façon dont les 2 groupes de patients sont constitués est sous le contrôle des expérimentateurs. Pour pouvoir établir des relations de causalité (entre administration du vaccin et guérison par exemple) ils doivent s'assurer que les groupes présentent les mêmes caractéristiques vis-à-vis de toutes les variables confondantes potentielles. Par exemple, on peut supposer que les hommes et les femmes ne réagissent pas de la même façon face au virus, ou face au vaccin. Ainsi, placer tous les hommes dans le premier groupe, et toutes les femmes dans le second groupe, serait évidemment une erreur. Car si le premier groupe guérit plus vite, comment peut-on être sûr de la cause de cette guérison ? Le premier groupe a-t-il guérit plus vite parce qu'il était constitué d'hommes, ou parce que les individus ont reçu le vaccin ? Puisque le sexe des individus est une variable confondante potentielle, il est important de répartir équitablement hommes et femmes dans les deux groupes. Et il en va de même pour énormément de variables confondantes potentielles : sexe des individus, âge, niveau d'études, revenu moyen, catégorie socio-professionnelle, etc. Ça n'est qu'en s'assurant que les 2 groupes sont homogènes vis-à-vis de l'ensemble de ces facteurs que les différences éventuelles qui seront observées à l'issue de l'expérience pourront être attribuées sans le moindre doute au traitement étudié : ici, l'administration du vaccin.

Dans le domaine de l'écologie, la plupart des études sont observationnelles, et elles permettent au mieux d'établir des corrélations, des liens entre variables, mais beaucoup plus rarement des liens de causalité formels. Il est toutefois souvent possible, après avoir observé un lien entre variables dans le milieu naturel, de mettre au point des expériences (donc des études expérimentales) permettant de tester des hypothèses précises, y compris des relations de causalité.

Par exemple, dans le milieu marin, en particulier littoral, l'apparition d'**imposex** chez certains mollusques² a pu être associé à la présence de tributylétain (TBT) à l'état de trace dans l'eau de mer> lorsque ce phénomène a été décelé, il était impossible d'affirmer que le TBT causait l'apparition d'imposex ; il n'y avait qu'une corrélation. Ça n'est que dans un second temps qu'une étude expérimentale en milieu contrôlé a permis d'établir un lien de causalité. Deux groupes de mollusques identiques en tous points sont placés dans différents bassins. On répartit ensuite de façon aléatoire les bassins en plusieurs lots, et chaque lot de bassin se voit attribuer un traitement : absence de TBT, TBT à la concentration X, TBT à la concentration Y, etc. C'est ce type d'étude expérimentale qui a permis d'établir avec certitude le caractère de perturbateur endocrinien du TBT, de prouver que l'imposex des mollusques pouvait être causé par le TBT, et de connaître les concentrations à partir desquelles les effets apparaissent.

² apparitions d'organes génitaux mâles chez des femelles saines par ailleurs

Important

Corrélation n'est pas causalité. Les études observationnelles ne peuvent (presque) jamais établir de lien de causalité formel entre variables. Au mieux, elles peuvent constater que des variables varient conjointement, dans le même sens ou en un sens opposé.

Seules les études expérimentales, dans lesquelles toutes les variables confondantes potentielles sont contrôlées, sont susceptibles de faire apparaître de véritables relations de cause à effet.

2.9 L'alternative non paramétrique

Quand les conditions d'application du test de corrélation de Pearson ne sont pas remplies (ce qui était le cas ici, voir Section 2.7.2), il faut faire un test équivalent non paramétrique. Le test utilisé le plus fréquemment dans cette situation est le test du ρ de Spearman (ρ est la lettre grecque “rho”, et non la lettre “p”). On l'effectue comme le test de Pearson en précisant simplement un argument supplémentaire : `method = "spearman"` (sans majuscule) :

```
cor.test(birds$NVisitsNestling, birds$futureBehavior,  
         method = "spearman")
```

```
Warning in cor.test.default(birds$NVisitsNestling, birds$futureBehavior, :  
Impossible de calculer la p-value exacte avec des ex-aequos
```

```
Spearman's rank correlation rho
```

```
data: birds$NVisitsNestling and birds$futureBehavior  
S = 1213.5, p-value = 0.01976  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
rho  
0.472374
```

Le test de Spearman est au test de Pearson ce que le test de Wilcoxon est au test de Student, ou ce que le test de Kruskal-Wallis est à l'ANOVA. Il travaille non pas sur les données brutes (ici, les mesures des scientifiques), mais sur des données modifiées, en l'occurrence, sur les rangs des données. La première conséquence évidente est une perte de puissance notable par rapport au test de Pearson. Cette perte de puissance peut être ici observée par le biais de la p -value plus élevée (donc moins significative) que pour le test précédent. Cela indique que même si la conclusion est la même, on rejette ici l'hypothèse nulle avec moins de confiance que pour le test de Pearson.

Le ρ de Spearman est équivalent au r de Pearson calculé sur les rangs des données. Lorsque plusieurs valeurs observées

sont égales, plusieurs valeurs ont le même rang, ce qui cause l'apparition du message d'avertissement suivant :

Impossible de calculer la p-value exacte avec des ex-aequos

Ce message est sans conséquence tant que la p -value du test de Spearman est éloignée du seuil α (ce qui est le cas ici). Mais quand $p \approx \alpha$, il faut être particulièrement prudent quant à l'interprétation qui est faite des résultats.

Enfin, comme pour le test de Pearson, il est possible de réaliser un test de Spearman unilatéral en utilisant l'argument `alternative = "less"` ou `alternative = "greater"`. Les précautions à prendre pour utiliser ce genre de test sont toujours les mêmes.

2.10 Exercices

2.10.1 *Canis lupus*

En 1970, le loup *canis lupus* a été éradiqué en Norvège et en Suède. Autour de 1980, un couple de loups, originaire d'une population plus à l'Est, a fondé une nouvelle population en Suède. En l'espace de 20 ans, cette population comptait approximativement 100 loups. Il y a toutefois fort à craindre qu'une population fondée par un si petit nombre d'individus souffre de consanguinité. Liberg et al. (2005) ont compilé les informations sur la reproduction dans cette population entre 1983 et 2002, et ils ont pu reconstruire le pédigrée des individus la composant. Ils ont ainsi été en mesure de déterminer avec précision le coefficient individuel de consanguinité dans 24 portées de louveteaux. Pour mémoire, le coefficient individuel de consanguinité vaut 0 si ses parents ne sont pas apparentés, 0.25 si ses parents sont frères et sœurs issus de grands-parents non apparentés, et plus de 0.25 si les associations consanguines se répètent depuis plusieurs générations.

On souhaite déterminer si le coefficient de consanguinité est associé à la probabilité de survie des jeunes durant leur premier hiver. Les données de Liberg et al. (2005) sont disponibles dans le fichier [loups.csv](#). La première colonne contient les coefficients de consanguinité et la seconde, le nombre de jeunes de chaque portée ayant survécu à leur premier hiver.

Vous analyserez ces données en suivant l'ordre des étapes décrites plus haut. En particulier, vous prendrez soin de :

- Vérifier la qualité des données.
- Mettre les données dans un format approprié si besoin.
- Réaliser une exploration statistique puis visuelle des données.
- Vérifier les conditions d'application d'un test paramétrique.
- Faire le test approprié en posant les hypothèses nulles et alternatives judicieuses.
- Répondre à la question posée en intégrant tous les éléments utiles.

2.10.2 Les miracles de la mémoire

À quel point les souvenirs d'événements miraculeux sont-ils fiables ? Une façon d'étudier cette question est de comparer différents récits de tours de magie extraordinaires. Parmi les tours célèbres, on trouve celui de la corde du fakir. Dans l'une de ses versions, un magicien jette l'extrémité d'une corde d'apparence normale en l'air et cette corde devient rigide. Un garçon grimpe à la corde et finit par disparaître en haut de la scène. Le magicien lui demande de répondre mais n'obtient pas de réponse. Il attrape alors un couteau, grimpe à son tour, et le garçon, découpé en morceaux, tombe du ciel dans un panier posé par terre. Le magicien redescend de la corde et aide le garçon vivant, en un seul morceau et non blessé, à sortir du panier.

Wiseman et Lamont (1996) ont retrouvé 21 récits écrits de ce tour par des personnes ayant elles-mêmes assisté à ce tour. Ils ont attribué un score à chaque description selon le caractère plus ou moins impressionnant de la description. Par exemple, un score de 1 était attribué si le récit faisait état que "le garçon grimpe à la corde, puis il en redescend". Les récits les plus impressionnants se sont vus attribuer la note de 5 ("le garçon grimpe, disparaît, est découpé en morceaux et réapparaît en chair et en os devant le public"). Pour chaque récit, les chercheurs ont également enregistré le nombre d'années écoulées entre le moment où le témoin a assisté au tour de magie, et le moment où il a consigné son récit par écrit.

Y a-t-il un lien entre le caractère impressionnant (“**impressiveness**”) d’un souvenir et le temps écoulé jusqu’à l’écriture de sa description (“**years**”) ? Si oui, cela pourrait indiquer une tendance de la mémoire humaine à exagérer et à perdre en précision avec le temps.

Les données de Wiseman et Lamont (1996) sont disponibles dans le fichier [ropetrick.csv](#). Importez ces données et analysez-les en respectant les consignes de l’exercice précédent.

3 Régression linéaire

3.1 Pré-requis

Comme pour chaque nouveau chapitre, je vous conseille de travailler dans un nouveau script que vous placerez dans votre répertoire de travail, et dans une nouvelle session de travail (Menu **Session** > **Restart R**). Inutile en revanche de créer un nouveau **Rproject** : vos pouvez tout à fait avoir plusieurs script dans le même répertoire de travail et pour un même **Rproject**. Comme toujours, consultez [le livre en ligne du semestre 3](#) si vous ne savez plus comment faire.

Si vous êtes dans une nouvelle session de travail (ou que vous avez quitté puis relancé **RStudio**), vous devrez penser à recharger en mémoire les packages utiles. Dans ce chapitre, vous aurez besoin d'utiliser :

- le **tidyverse** (Wickham 2023), qui comprend notamment le package **readr** (Wickham, Hester, et Bryan 2023), pour importer facilement des fichiers **.csv** au format **tibble**, le package **dplyr** (Wickham, François, et al. 2023), pour manipuler des tableaux, et le package **ggplot2** (Wickham, Chang, et al. 2023) pour les représentations graphiques.
- **skimr** (Waring et al. 2022), qui permet de calculer des résumés de données très informatifs.
- **datasauRus** (Davies, Locke, et D'Agostino McGowan 2022), qui fournit plusieurs jeux de données fictifs que nous examinerons en guise d'exercices.

```
library(tidyverse)
library(skimr)
library(datasauRus)
```

Vous aurez également besoin des jeux de données suivants, qu'il vous faut donc télécharger dans votre répertoire de travail :

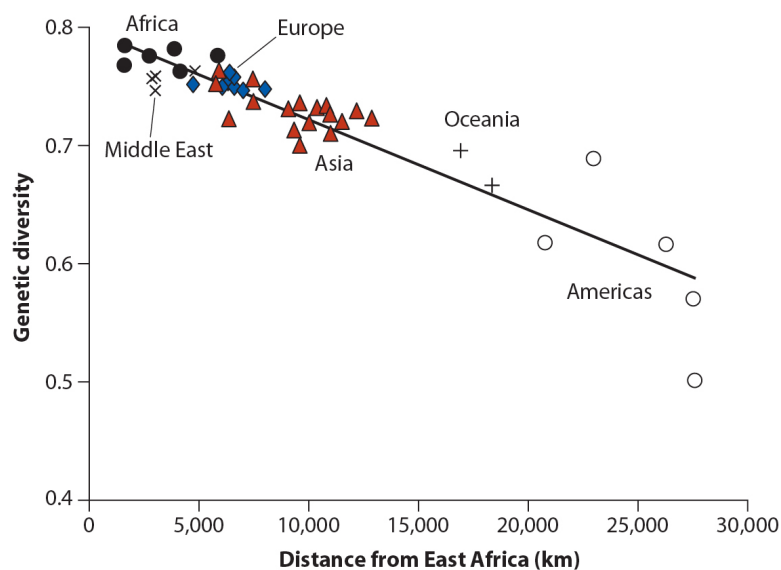
- `plantbiomass.csv`
- `hockey.csv`

```
theme_set(theme_bw())
```

3.2 Principe

La **régression linéaire** est une méthode qui fait partie de la famille des **modèles linéaires**, tout comme l'**ANOVA**. ANOVA et régression linéaires sont en effet deux méthodes très proches, et leur mise en œuvre dans **RStudio** présente de nombreuses similitudes, tant dans la syntaxe des fonctions que nous utiliserons, que dans la façon de vérifier les conditions d'application.

La régression linéaire est une méthode souvent utilisée pour prédire les valeurs d'une variable numérique (appelée variable expliquée) à partir des valeurs d'une seconde variable (appelée variable explicative). Par exemple, le nuage de points de la figure ci-dessous montre comment la diversité génétique dans une population humaine locale peut être prédite par sa distance de dispersion depuis l'Est africain en ajustant une droite aux données (d'après Whitlock et Schluter 2015). L'homme moderne est apparu en Afrique et nos ancêtres ont perdu un peu de diversité génétique à chaque étape de leur colonisation de nouveaux territoires.



Contrairement à la corrélation, ici, on n'examine pas seulement une éventuelle liaison entre 2 variables numériques : on suppose qu'une variable peut être (en partie) expliquée par une autre. Nous aurons donc à distinguer les variables expliquées (ou dépendantes) qui figureront sur l'axe des ordonnées et seront nos variables prédites, et les variables explicatives (ou indépendantes) qui figureront sur l'axe des abscisses et seront les prédictors.

Contrairement à la corrélation qui, comme nous l'avons expliqué en détail, ne permet pas d'aborder les questions de causalité, lorsque l'on s'intéresse à la régression linéaire, on essaie au contraire de prédire ou d'expliquer les variations de la variable expliquée par celles de la variable explicative. En d'autres termes, on considère que les variations de la variable explicative sont au moins en partie la cause des variations de la variable expliquée.

Lorsque l'on s'intéresse à la régression linéaire, on considère que la relation qui lie les deux variables est linéaire, et on souhaite **quantifier l'intensité de cette relation** (quand la variable explicative augmente d'une unité, de combien d'unité la variable expliquée augmente ou diminue-t-elle ?). Nous allons voir maintenant comment mettre en œuvre cette méthode dans RStudio.

3.3 Contexte

Les activités humaines réduisent le nombre d'espèces dans un grand nombre d'écosystèmes à la surface du globe. Est-ce que cette diminution du nombre d'espèces affecte le fonctionnement de base des écosystèmes ? Où est-ce qu'au contraire, les espèces végétales sont majoritairement interchangeables, les fonctions écologiques des espèces disparues³ pouvant être assurées par les espèces toujours présentes ?

³ par exemple, la production d'O₂ et la fixation de CO₂, la dépollution des sols, leur fixation, la protection contre les inondations et l'érosion...

Pour tenter de répondre à cette question, Tilman, Reich, et Knops (2006) ont ensemencé 161 parcelles de 9 mètres sur 9 mètres dans la réserve de Cedar Creek (Minnesota, USA). Ils ont utilisé un nombre variable d'espèces typiques des prairies et ont mesuré la production de biomasse de chaque parcelle pendant 10 ans. Des lots de 1, 2, 4, 8 ou 16 plantes pluri-annuelles (choisies au hasard parmi une liste de 18 espèces possibles) ont été assignés au hasard dans chacune des 161 parcelles. À l'issue des 10 années d'étude, les chercheurs ont mesuré un indice de stabilité de la biomasse en divisant la moyenne des biomasses sur 10 ans, par l'écart-type de ces mêmes biomasses.

3.4 Importation et mise en forme des données

Les données de cette expérience sont disponibles dans le fichier [plantbiomass.csv](#).

Comme toujours, on importe les données et on commence par un examen visuel afin de détecter les éventuels problèmes et pour savoir où l'on va.

```
plant
```

```
# A tibble: 161 x 2
  nSpecies biomassStability
  <dbl>         <dbl>
1       1         2.01
2       1         1.91
3       1         1.89
```

```

4      1      1.86
5      1      1.74
6      1      1.66
7      1      1.57
8      1      1.48
9      1      1.48
10     1      1.45
# i 151 more rows

```

Ce premier examen nous montre que nous disposons bien de 161 observations pour 2 variables : le nombre d'espèces présentes dans la parcelle pendant 10 ans, et l'indice de stabilité de la biomasse de chaque parcelle. Visiblement, les données sont au bon format, on dispose bien de toutes les variables dont on a besoin et leurs noms sont parlants. Nous n'aurons donc pas besoin de modifier quoi que ce soit dans ces données.

3.5 Exploration statistique des données

Comme toujours, on examine quelques statistiques descriptives de position et de dispersion (voir d'incertitude), pour se faire une idée de la forme des données et pour repérer les éventuelles données manquantes ou valeurs aberrantes.

```

skim(plant)

-- Data Summary -----
Name                Values
Number of rows      plant
Number of columns    161
                    2
-----
Column type frequency:
  numeric            2
-----
Group variables      None

-- Variable type: numeric -----
  skim_variable  n_missing complete_rate mean    sd  p0  p25  p50  p75  p100
1 nSpecies      0           1 6.32 5.64  1    2    4    8   16

```



```
2 biomassStability      0          1 1.41 0.394 0.293 1.12 1.39 1.65 2.76
  hist
1
2
```

Ce premier examen nous montre que nous n'avons aucune données manquantes et que l'indice de stabilité a une distribution à peu près symétrique et qu'il varie d'un peu plus de 0.3 à près de 2.8. Pour en apprendre un peu plus, nous pouvons examiner les données en groupes. Ici, la variable `nSpecies` est bien une variable numérique, mais elle prend seulement quelques valeurs entières (1, 2, 4, 8 ou 16 espèces). Il est donc possible de regarder les valeurs de stabilité de biomasse pour chaque nombre d'espèces dans les parcelles:

```
plant %>%
  group_by(nSpecies) %>%
  skim()
```

```
-- Data Summary -----
```

	Values
Name	Piped data
Number of rows	161
Number of columns	2

```
-----
Column type frequency:
```

numeric	1
---------	---

```
-----
Group variables      nSpecies
```

```
-- Variable type: numeric -----
```

	skim_variable	nSpecies	n_missing	complete_rate	mean	sd	p0	p25	p50
1	biomassStability	1	0	1	1.21	0.388	0.728	0.880	1.10
2	biomassStability	2	0	1	1.28	0.360	0.293	1.09	1.32
3	biomassStability	4	0	1	1.31	0.314	0.756	1.10	1.35
4	biomassStability	8	0	1	1.50	0.251	0.928	1.32	1.48
5	biomassStability	16	0	1	1.71	0.403	1.08	1.39	1.66

p75 p100 hist

1	1.48	2.01
2	1.51	2.00
3	1.51	2.00
4	1.61	2.05
5	1.96	2.76

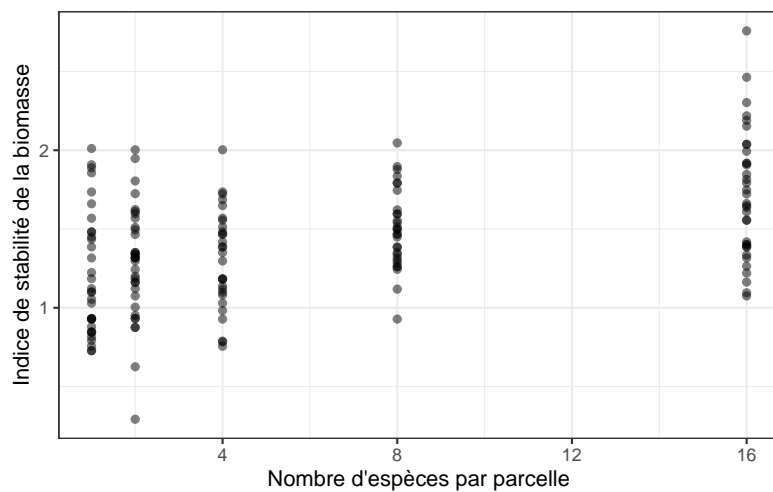
Cette fois, on obtient des informations pour chaque groupe de parcelles contenant un nombre d'espèces spécifique. On constate par exemple que la moyenne de l'indice de stabilité de la biomasse augmente très peu entre les catégories 1, 2 et 4 espèces par parcelle, mais que l'augmentation semble plus marquée pour 8 et 16 espèces par parcelle. Tous les écarts-types semblent très proches. Les parcelles avec 1 et 16 espèces en particulier présentent des histogrammes nettement asymétriques.

Comme pour la corrélation, il est inutile ici de calculer des indices d'imprécision. Ça n'est pas la moyenne de ces variables qui nous intéresse, mais la relation entre elles. Nous serons en revanche amenés à calculer des intervalles de confiance à 95% des paramètres de la régression linéaire, puisque ce sont eux qui nous permettront de qualifier (et quantifier) la relation entre les 2 variables.

3.6 Exploration graphique des données

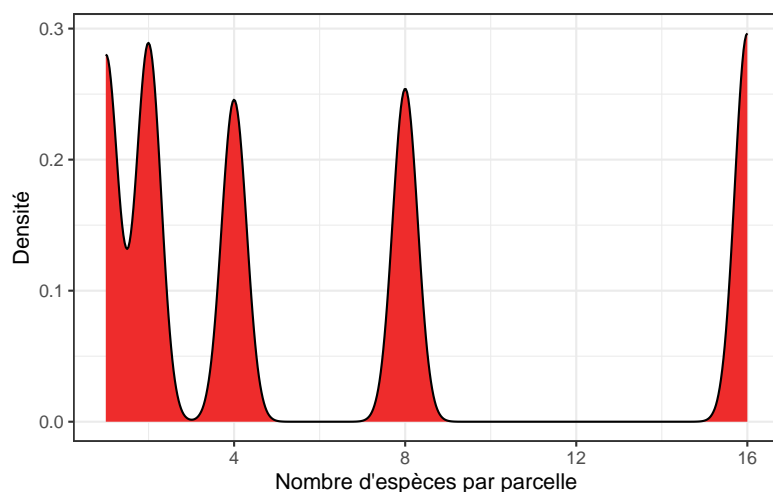
Visualiser les données est toujours aussi indispensable. Ici, comme pour la corrélation, on commence par un nuage de points pour visualiser les données et la forme de leur relation :

```
plant %>%  
  ggplot(aes(x = nSpecies, y = biomassStability)) +  
  geom_point(alpha = 0.5) +  
  labs(x = "Nombre d'espèces par parcelle",  
       y = "Indice de stabilité de la biomasse")
```



Ce graphique nous apprend que contrairement à la plupart des méthodes statistiques vues jusqu'ici, il n'est pas nécessaire que les données des variables soient distribuées selon une loi Normale. Ici, nous avons des données qui sont tout sauf normales pour la variable explicative puisque nous avons seulement les entiers 1, 2, 4, 8 et 16. Un histogramme ou une courbe de densité montre que la distribution de cette variable est très loin de la Normalité :

```
plant %>%
  ggplot(aes(x = nSpecies)) +
    geom_density(fill = "firebrick2", adjust = 0.2) +
    labs(x = "Nombre d'espèces par parcelle",
         y = "Densité")
```

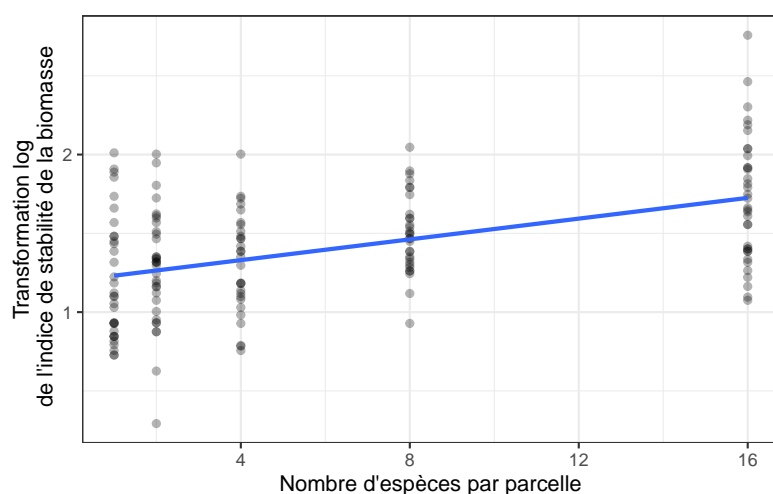


Cela n'est pas du tout problématique : comme pour l'ANOVA, les conditions d'application porteront sur les **résidus de la régression**, pas sur les variables elles-mêmes. Comme pour l'ANOVA, ce sont les résidus de la régression qui devront suivre une distribution Normale, pas les variables de départ.

On peut visualiser dès maintenant la droite de régression linéaire qui permet de lier ces deux variables grâce à la fonction `geom_smooth(method = "lm", se = FALSE)` :

```
plant %>%
  ggplot(aes(x = nSpecies, y = biomassStability)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Nombre d'espèces par parcelle",
       y = "Transformation log\n de l'indice de stabilité de la biomasse")
```

``geom_smooth()`` using formula = 'y ~ x'



L'argument `method = "lm"` indique qu'on souhaite ajouter une droite de régression sur le graphique, et `se = FALSE` permet de ne faire apparaître que la droite de régression, sans son intervalle d'incertitude. Nous reviendrons sur la notion d'incertitude de la régression un peu plus loin.

À supposer que nous ayons le droit d'effectuer une régression linéaire (ce qu'il faudra vérifier avec les conditions

d'application, **après** avoir fait la régression), la pente devrait être positive.

3.7 Le test paramétrique

3.7.1 Les hypothèses

À une exception près, la procédure de régression linéaire est en tous points identique à l'analyse de variance. Quand on fait une ANOVA, la variable expliquée est numérique et la variable explicative est catégorielle (c'est un facteur). Dans **RStudio**, la formule ressemble donc à ceci :

$$Y \sim F$$

Quand on fait une régression linéaire, les 2 variables sont numériques. La formule dans 'RStudio ressemble donc à ça :

$$Y \sim X$$

Dans ces formules, **Y** est la variable numérique expliquée, **F** est une variable catégorielle (ou facteur) et **X** est une variable numérique explicative. La forme est donc très proche, et tout le reste est identique : on exprime la variable expliquée en fonction de la variable explicative et on vérifie après coup, grâce aux résidus, si nous avons le droit ou non de faire l'analyse.

La différence majeure entre ANOVA et régression linéaire concerne les hypothèses du test. Faire une régression linéaire revient en effet à effectuer en même temps 2 tests d'hypothèses indépendants : le premier concerne l'ordonnée à l'origine de la droite de régression et le second concerne la pente de la droite de régression. On ne parle donc plus de comparer des moyennes entre groupes : on cherche à déterminer si la pente et l'ordonnée à l'origine de la meilleure droite de régression possible valent zéro ou non. Les hypothèses de ces tests sont les suivantes :

Pour l'ordonnée à l'origine ("intercept" en anglais) :

- H_0 : l'ordonnée à l'origine de la droite de régression vaut 0 dans la population générale.
- H_1 : l'ordonnée à l'origine de la droite de régression est différente de 0 dans la population générale.

Pour la pente (“slope” en anglais) :

- H_0 : la pente de la droite de régression vaut 0 dans la population générale. Autrement dit, il n’y a pas de lien entre les deux variables.
- H_1 : la pente de la droite de régression est différente de 0 dans la population générale. Autrement dit, il y a bien un lien entre les deux variables étudiées.

Vous notez qu’ici, comme pour tous les autres tests statistiques traités dans ce livre en ligne, les tests ne permettent que de rejeter ou non les hypothèses nulles. Si on rejette ces hypothèses, le test ne nous dit rien de la valeur de la pente et de l’ordonnée à l’origine. On sait que ces paramètres sont significativement différents de zéro, mais rien de plus. Il faudra alors recourir à l’estimation pour déterminer la valeur de ces paramètres, ainsi que leurs intervalles d’incertitude.

3.7.2 Réalisation du test

Pour faire une régression linéaire dans **RStudio**, on utilise la fonction `lm()` (comme **linear model**). Et comme pour l’ANOVA, les résultats de l’analyse doivent être stockés dans un objet puisque cet objet contiendra tous les éléments utiles pour vérifier les conditions d’application :

```
reg1 <- lm(biomassStability ~ nSpecies, data = plant)
```

Comme pour l’ANOVA, on affiche les résultats de ces tests à l’aide de la fonction `summary()`

```
summary(reg1)
```

Call:

```
lm(formula = biomassStability ~ nSpecies, data = plant)
```

```

Residuals:
      Min       1Q   Median       3Q      Max
-0.97148 -0.25984 -0.00234  0.23100  1.03237

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.198294    0.041298  29.016 < 2e-16 ***
nSpecies     0.032926    0.004884   6.742 2.73e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3484 on 159 degrees of freedom
Multiple R-squared:  0.2223,    Adjusted R-squared:  0.2174
F-statistic: 45.45 on 1 and 159 DF,  p-value: 2.733e-10

```

Dans la forme, ces résultats sont très proches de ceux de l'ANOVA. La rubrique **Residuals** donne des informations sommaires sur les résidus. Ces informations sont utiles puisque les résidus serviront à vérifier les conditions d'application de la régression. À ce stade, on regarde surtout si la médiane des résidus est proche de 0 et si les résidus sont à peu près symétriques (les premier et troisième quartiles ont à peu près la même valeur absolue, idem pour le minimum et le maximum).

Le tableau **Coefficients** est celui qui nous intéresse le plus puisqu'il nous fournit, outre la réponse aux 2 tests, les estimations pour l'ordonnée à l'origine et la pente de la droite de régression.

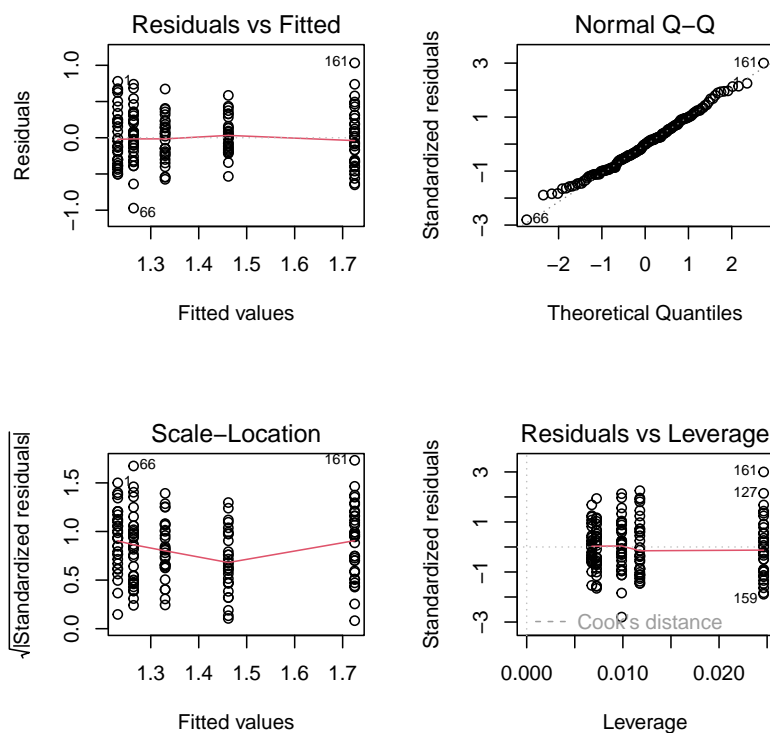
Avant d'aller plus loin dans l'interprétation de ces résultats, il nous faut déterminer si nous avons bel et bien le droit de réaliser cette régression, en vérifiant si ses conditions d'application sont remplies.

3.7.3 Conditions d'application

Les conditions d'application de la régression sont les mêmes que celles de l'ANOVA. Je vous renvoie donc à la Section 1.6.2 pour savoir quelles sont ces conditions d'application et comment les vérifier. J'insiste bien sur le fait que les conditions d'application sont absolument identiques à celles de l'ANOVA. Si je fais ici l'économie de la description,

vous ne devez **jamais faire l'économie** de la vérification des conditions d'application.

```
par(mfrow = c(2, 2))
plot(reg1)
```



```
par(mfrow = c(1, 1))
```

C'est seulement après avoir réalisé, examiné et commenté ces graphiques que vous serez en mesure de dire si oui ou non vous aviez le droit de faire la régression linéaire, et donc d'en interpréter les résultats.

Ici, les conditions d'application semblent tout à fait remplies :

1. Les deux graphiques de gauche confirment que les résidus sont homogènes. En particulier, sur le premier graphique (en haut à gauche), la ligne rouge est presque parfaitement horizontale, il y a à peu près autant de résidus au-dessus qu'en dessous de la ligne pointillée, et les résidus pourraient rentrer dans une boîte ayant

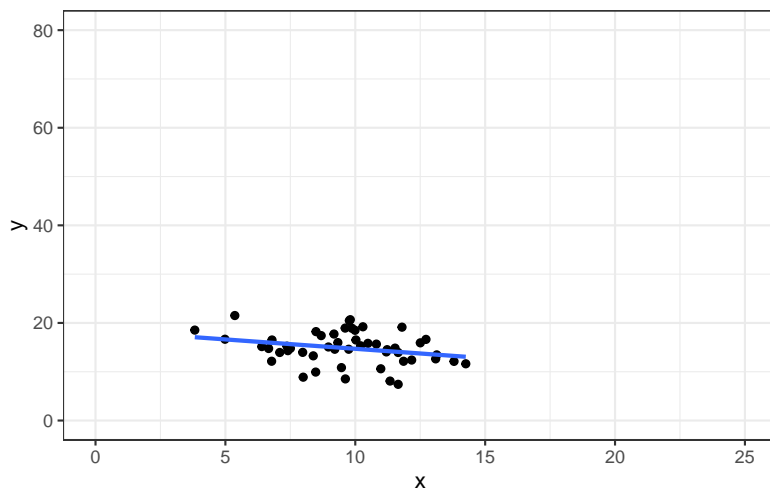
la même hauteur d'un bout à l'autre du graphique (pas d'effet "entonnoir" ou "nœud papillon").

2. Le graphique quantile-quantile (en haut à droite), montre des points qui sont presque parfaitement alignés sur la droite pointillée, indiquant des résidus distribués selon une distribution Normale.

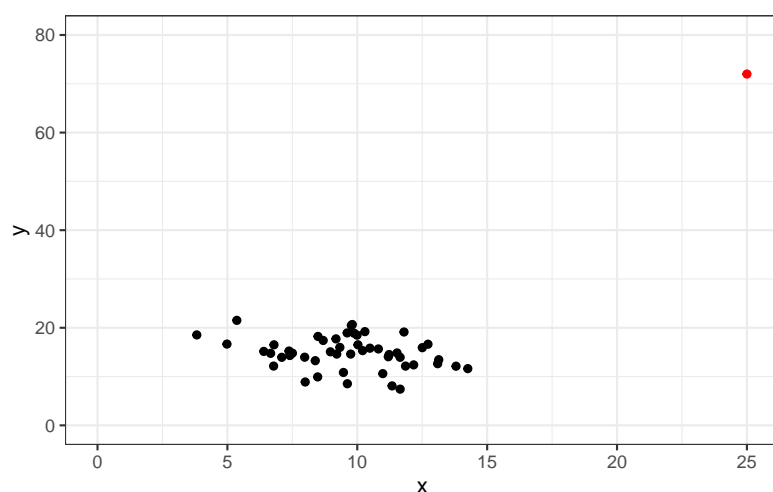
On pourrait vérifier ces éléments avec des tests statistiques (encore une fois, reportez vous à la Section 1.6.2 si vous ne vous rappelez plus comment faire), mais c'est ici inutile tant les conditions semblent bien respectées.

Le dernier graphique (en bas à droite, "Residuals vs Leverage") ne permet pas de vérifier les conditions d'application à proprement parler, mais permet de repérer des points ayant un poids trop important dans l'analyse. Ces points devraient être retirés s'il y en a (ce qui n'est pas le cas ici), car leur influence est tellement forte qu'ils faussent grandement les résultats de l'analyse. Pour voir à quoi ce graphique ressemble quand de tels points sont présents, je représente ci-dessous un exemple fictif.

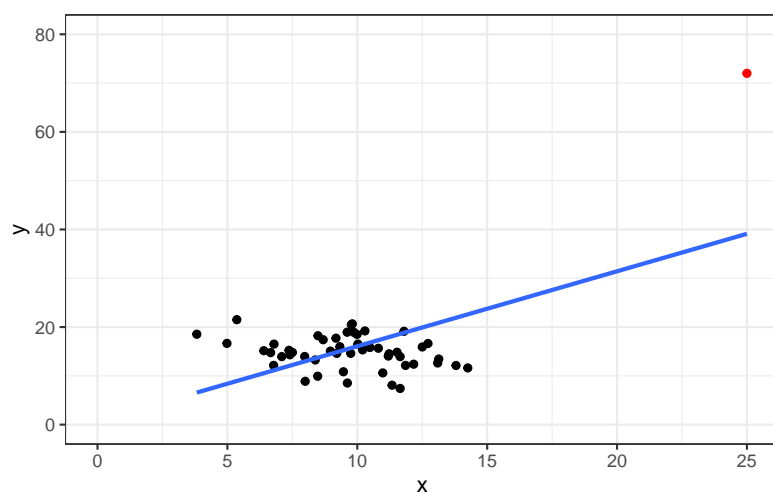
Imaginez un jeu de données dans lequel absolument aucune tendance n'est présente. Le nuage de points d'un tel jeu de données devrait être approximativement circulaire, avec une droite de régression presque horizontale, indiquant une absence de lien entre les 2 variables étudiées x et y :



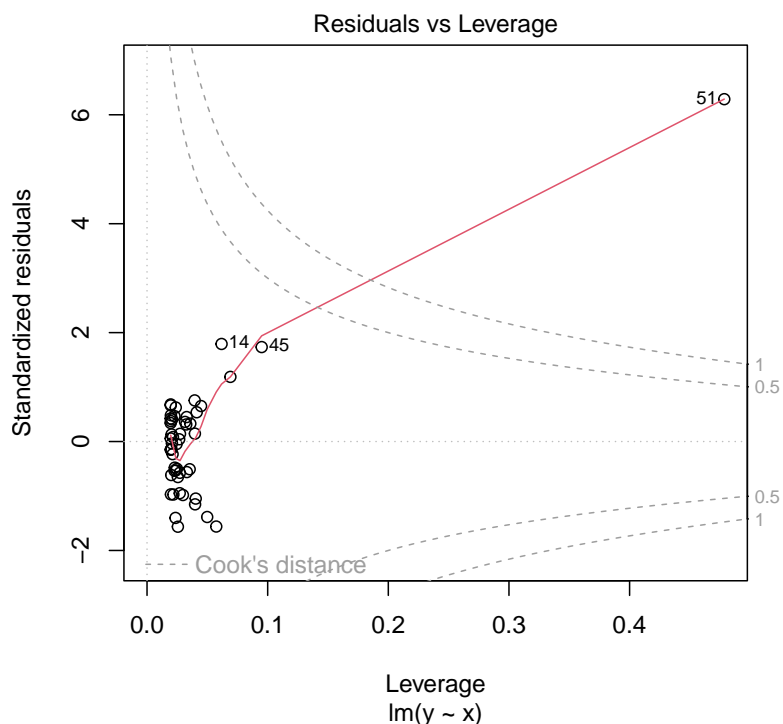
Imaginez maintenant qu'on ajoute à ces données un unique point (en rouge sur le graphique), très éloigné des autres :



La seule présence de ce point modifierait très fortement les résultats de la régression linéaire :



Sans ce point supplémentaire, la droite de régression a une pente légèrement négative, avec ce point, la pente est fortement positive. Il n'est pas normal qu'une observation unique prenne le pas sur toutes les autres (il y en a 50) et qu'elle affecte autant les résultats de la régression. la situation est ici caricaturale, et on voit bien qu'il faudrait retirer la valeur atypique pour obtenir des résultats censés. Les points ayant une influence démesurée sur les résultats ne sont pas toujours aussi évidents à repérer. C'est justement à cela que sert le graphique "Residuals vs Leverage" :



Sur ce graphique, les points qui apparaissent au-delà des lignes pointillées (en haut à droite ou en bas à gauche du graphique) sont ceux qui ont une influence trop forte sur les résultats et qu'il faudrait donc retirer des données pour obtenir des résultats plus représentatifs de la tendance observée pour la majorité des points.

Si je reviens à nos données de stabilité des biomasses en fonction du nombre d'espèces par parcelles, les lignes courbes pointillées qui délimitent les zones "à problème" ne sont même pas visibles sur le graphique. Nous n'avons donc pas de points problématiques.

Au final, les conditions d'application de la régression sont parfaitement vérifiées et nous pouvons donc en interpréter les résultats.

3.7.4 Interprétation des résultats

Revenons donc à l'affichage des résultats :

```
summary(reg1)
```

```

Call:
lm(formula = biomassStability ~ nSpecies, data = plant)

Residuals:
    Min       1Q   Median       3Q      Max
-0.97148 -0.25984 -0.00234  0.23100  1.03237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.198294   0.041298  29.016 < 2e-16 ***
nSpecies      0.032926   0.004884   6.742 2.73e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3484 on 159 degrees of freedom
Multiple R-squared:  0.2223,    Adjusted R-squared:  0.2174
F-statistic: 45.45 on 1 and 159 DF,  p-value: 2.733e-10

```

Outre une description synthétique de la distribution des résidus, ces résultats nous apprennent que :

- l'ordonnée à l'origine (intercept) est estimée à 1.198 (rappelez-vous que cette valeur fait référence à l'indice de stabilité de la biomasse)
- la pente est estimée à 0.033 (quand le nombre d'espèces augmente d'une unité, l'indice de stabilité de la biomasse augmente de 0.033 unités)

Les p -values de chacun des 2 tests sont fournies dans la dernière colonne et sont ici très inférieures à α : on rejette donc les 2 hypothèses nulles. En particulier, puisque l'hypothèse nulle est rejetée pour le test qui concerne la pente de la droite, on peut considérer que le nombre de plantes dans les parcelles influence bel et bien l'indice de stabilité de la biomasse. Autrement dit, le nombre de plantes dans les parcelles, permet, dans une certaine mesure, de prédire la valeur de l'indice de stabilité de la biomasse.

La relation n'est toutefois pas très forte : le nombre de plantes dans chaque parcelle ne permet de prédire l'indice de stabilité de la biomasse que dans une mesure assez faible. C'est le **Adjusted R-squared** qui nous indique quelle est la "qualité" de prédiction du modèle. Ici, il vaut 0.22. Cela signifie

que 22% des variations de l'indice de stabilité de la biomasse sont prédits par le nombre de plantes dans les parcelles. Une autre façon de présenter les choses consiste à dire que 78% des variations de l'indice de stabilité de biomasse sont expliqués par d'autres facteurs que celui que nous avons pris en compte dans notre modèle de régression linéaire (*i.e.* le nombre d'espèces par parcelle). Le R^2 (à ne pas confondre avec le coefficient de corrélation r) renseigne sur la qualité de l'ajustement des données à la droite de régression. Il nous indique ici que le pouvoir prédictif de notre modèle linéaire est assez faible. Il est néanmoins significatif, ce qui indique que notre variable explicative joue bel et bien un rôle non négligeable dans les variations de la variable expliquée. Une autre façon de comprendre ce résultat est la suivante : si on connaît le nombre de plantes dans une parcelle, on peut prédire 22% de la valeur de l'indice de stabilité de la biomasse.

3.7.5 Intervalle de confiance de la régression

L'équation de notre droite de régression vaut donc :

$$y = 0.033 \times x + 1.198$$

Avec y , l'indice de stabilité de la biomasse, et x , le nombre d'espèces par parcelles. On voit bien que la droite nous permet de prédire une valeur d'indice de stabilité de la biomasse pour un nombre d'espèces donnée par parcelle, y compris pour des nombres d'espèces qui n'ont pas été testés. par exemple, pour $n = 6$ espèces, on peut s'attendre à un indice de stabilité de la biomasse de $0.033 \times 6 + 1.198 = 1.396$. Il convient toutefois de prendre deux précautions très importantes quand on fait ce genre prédiction :

1. la régression et son équation ne sont valables que sur l'intervalle que nous avons étudié pour la variable explicative. Ainsi, on peut faire des prédictions pour des valeurs de nombre d'espèces comprises entre 1 et 16, mais pas au-delà. En effet, rien ne nous dit que cette relation reste valable au-delà de la gamme $n = [1 ; 16]$. Peut-être la relation change-t-elle de nature à partir de $n = 20$ espèces par parcelles. Peut-être que la pente devient nulle ou négative. Ou peut-être la relation n'est-elle plus linéaire au-delà de $n = 16$ espèces par parcelle.

En bref, puisque nous n'avons des informations sur le comportement de notre système d'étude que pour une gamme de valeurs précises sur l'axe des x , il nous est impossible de prédire quoi que ce soit en dehors de cette gamme de valeurs.

2. Même à l'intérieur de la gamme de valeur permettant de faire des prédictions, toute prédiction est entachée d'incertitude. Pour s'en convaincre, il suffit de regarder la grande dispersion des valeurs observées pour y pour chaque valeur de x . Par exemple, pour $x = 8$ espèces par parcelle, les valeurs observées pour l'indice de stabilité vont de moins de 1 à plus de 2. Le modèle prédit une valeur d'environ 1.46 ($0.033 * 8 + 1.198 = 1.462$), mais on voit bien que l'incertitude persiste. C'est la raison pour laquelle on aura toujours besoin de calculer des **indices d'incertitude**, pour avoir une idée de l'erreur commise lorsque l'on fait une prédiction.

Prudence avec les prédictions

Une droite de régression permet de faire des prédictions :

- uniquement sur la gamme de valeurs de l'axe des x qui a permis d'établir l'équation de la droite de régression
- avec une imprécision/incertitude qu'il est toujours nécessaire d'estimer.

La pente et l'ordonnée à l'origine de cette droite de régression ont été obtenues à partir des données d'un échantillon (ici, $n = 161$ parcelles). Il s'agit donc d'estimations des pentes et ordonnées à l'origine de la relation plus générale qui concerne la population globale, mais que nous ne connaissons jamais avec précision. Comme toute estimation, les valeurs de pente et d'ordonnée à l'origine de la droite de régression sont entachées d'**incertitude**. Nous pouvons quantifier ces incertitudes grâce au calcul des intervalles de confiance à 95% de ces 2 paramètres :

```
confint(reg1)
```

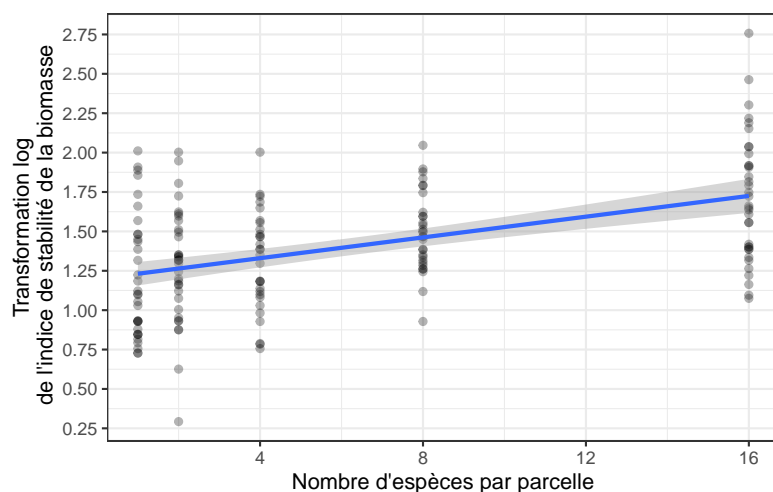
	2.5 %	97.5 %
(Intercept)	1.11673087	1.27985782
nSpecies	0.02328063	0.04257117

Ces résultats nous indiquent que les valeurs d'ordonnées à l'origine les plus probables dans la population générale sont vraisemblablement comprises entre 1.117 et 1.280. De même, les valeurs de pentes les plus probables dans la population générale sont vraisemblablement situées dans l'intervalle [0.023 ; 0.043]. Autrement dit, pour la pente de la droite de régression, la meilleure estimation possible vaut 0.033, mais dans la population générale, les valeurs comprises dans l'intervalle [0.023 ; 0.043] sont parmi les plus probables.

Il est possible de visualiser cette incertitude grâce à la fonction `geom_smooth()` utilisée plus tôt, en spécifiant `se = TRUE` :

```
plant %>%
  ggplot(aes(x = nSpecies, y = biomassStability)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(x = "Nombre d'espèces par parcelle",
       y = "Transformation log\n de l'indice de stabilité de la biomasse") +
  scale_y_continuous(breaks = seq(0, 3, 0.25))
```

``geom_smooth()`` using formula = 'y ~ x'



Dans la population générale, la vraie droite de régression peut se trouver n'importe où dans la bande grise. Cet intervalle d'incertitude est bien moins large que l'étendue des données sur l'axe des ordonnées, et c'est tant mieux. Il correspond à l'**incertitude de la moyenne** de l'indice de stabilité de la biomasse pour une valeur donnée de la variable explicative. Ainsi, par exemple, si on réalise une expérience avec plusieurs parcelles contenant toutes 8 espèces, alors, la régression et son incertitude associée nous disent que **la moyenne** de l'indice de stabilité de la biomasse vaudra environ 1.46 (la valeur de la droite de régression pour $x = 8$), avec un intervalle de confiance à 95% de [1.40 ; 1.51] (l'étendue de la zone grisée autour de la courbe pour $n = 8$ espèces).

3.8 L'alternative non paramétrique

Lorsque les conditions d'application de la régression linéaire ne sont pas vérifiées, on a principalement deux options :

1. On essaie de transformer les données afin que les résidus de la régression se comportent mieux. Cela signifie tester différents types de transformations (passage au logarithme, à l'inverse, à la racine carrée...), ce qui peut être chronophage pour un résultat pas toujours garanti. Il existe de très nombreuses transformations et trouver la meilleure n'est pas trivial. Par ailleurs, l'interprétation d'une relation linéaire impliquant des données transformées n'est pas toujours aisée.
2. On utilise d'autres types de modèles de régression, en particulier les modèles de régressions linéaires généralisées (GLM), qui s'accommodent très bien de résidus non normaux et/ou non homogènes. Ces méthodes sont aujourd'hui préférées à la transformation des données et elles donnent de très bons résultats. Mais il s'agit là d'une toute autre classe de méthodes qui ne sont pas au programme de la licence.

3.9 Exercices

3.9.1 Datasaurus et Anscombe

Exécutez les commandes suivantes :

```
library(datasauRus)

datasaurus_dozen %>%
  group_by(dataset) %>%
  summarize(
    moy_x      = mean(x),
    moy_y      = mean(y),
    ecart_type_x = sd(x),
    ecart_type_y = sd(y),
    correl_x_y  = cor(x, y),
    pente      = coef(lm(y~x))[2],
    ordonnee_origine = coef(lm(y~x))[1]
  )
```

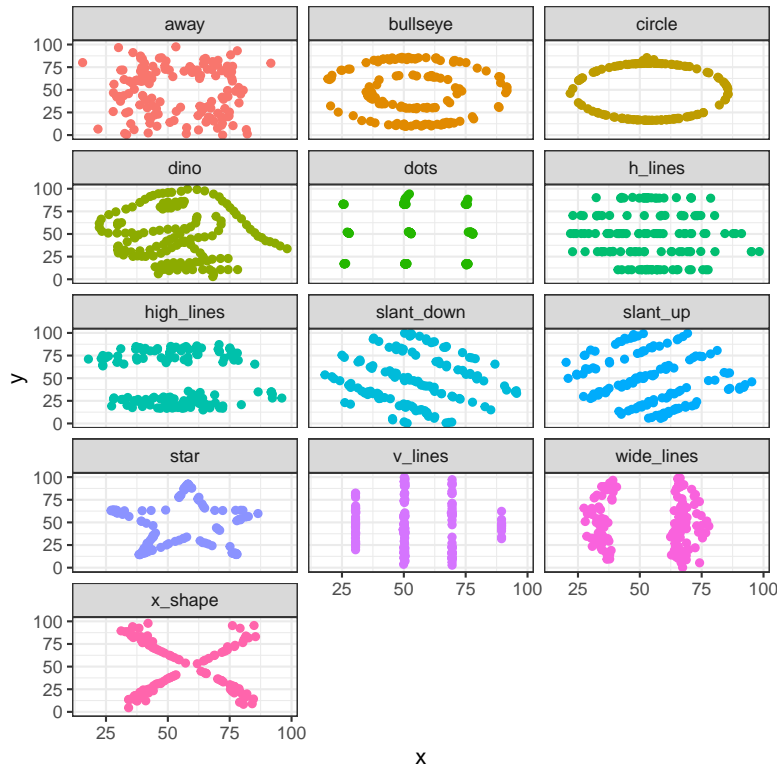
Examinez attentivement les nombreux résultats produits par cette commande. Vous devriez remarquer que pour ces 13 jeux de données, 2 variables numériques x et y sont mises en relation. Pour tous ces jeux de données, on observe que la moyenne de tous les x est la même, la moyenne de tous les y est la même, les écarts-types des x sont identiques, les écarts-types des y aussi, la corrélation entre x et y est également extrêmement proche pour tous les jeux de données, et lorsque l'on effectue une régression linéaire de y en fonction de x , les ordonnées à l'origine et les pentes des droites de régression sont extrêmement proches pour les 13 jeux de données.

Si on s'en tient à ces calculs d'indices synthétiques, on pourrait croire que ces jeux de données sont identiques ou presque. Pourtant, ce n'est pas par hasard que je vous répète à longueur de temps qu'il est **indispensable de regarder les données** avant de se lancer dans les analyses et les statistiques. Car ici, ces jeux de données sont très différents ! Conclure qu'ils sont identiques simplement parce que les statistiques descriptives sont égales, serait une erreur majeure :

```

datasaurus_dozen %>%
  ggplot(aes(x, y, color = dataset)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(~dataset, ncol = 3) +
  theme_bw()

```



Le quartet d'Anscombe est un autre exemple de ce type de problème.

Dans la console, exécutez la commande suivante (il vous faudra peut-être presser la touche Entrée plusieurs fois) pour produire les 4 graphiques d'Anscombe :

```
example(anscombe)
```

Examinez attentivement les nombreux résultats produits par cette commande dans la console, ainsi que les 4 graphiques obtenus. Vous devriez remarquer que pour ces 4 jeux de données, 2 variables numériques sont là encore mises en relation, et qu'elles présentent toutes les mêmes caractéristiques. En particulier, les régressions linéaires ont toutes les mêmes pentes

et ordonnées à l'origine. Pourtant, seule l'une de ces régressions linéaires est valide. Pourquoi ?

3.9.2 In your face

Les hommes ont en moyenne un ratio “largeur du visage sur longueur du visage” supérieur à celui des femmes. Cela reflète des niveaux d'expression de la testostérone différents entre hommes et femmes au moment de la puberté. On sait aussi que les niveaux de testostérone permettent de prédire, dans une certaine mesure, l'agressivité chez les mâles de nombreuses espèces. On peut donc poser la question suivante : la forme du visage permet-elle de prédire l'agressivité ?

Pour tester cela, Carré et McCormick (2008) ont suivi 21 joueurs de hockey sur glace au niveau universitaire. Ils ont tout d'abord mesuré le ratio largeur du visage sur longueur du visage de chaque sujet, puis, ils ont compté le nombre moyen de minutes de pénalité par match reçu par chaque sujet au cours de la saison, en se limitant aux pénalités infligées pour cause de brutalité. Les données sont fournies dans le fichier [hockey.csv](#).

Importez, examinez et analysez ces données pour répondre à la question posée.

4 Comparaison de proportions

Ne sera finalement pas traité cette année faute de temps...

References

- Campbell, Scott S., et Patricia J. Murphy. 1998. « Extraocular Circadian Phototransduction in Humans ». *Science* 279 (5349): 396-99. <https://doi.org/10.1126/science.279.5349.396>.
- Carré, Justin M, et Cheryl M McCormick. 2008. « In Your Face: Facial Metrics Predict Aggressive Behaviour in the Laboratory and in Varsity and Professional Hockey Players ». *Proceedings of the Royal Society B: Biological Sciences* 275 (1651): 2651-56. <https://doi.org/10.1098/rspb.2008.0873>.
- Davies, Rhian, Steph Locke, et Lucy D'Agostino McGowan. 2022. *datasauRus: Datasets from the Datasaurus Dozen*. <https://CRAN.R-project.org/package=datasauRus>.
- Fox, John, Sanford Weisberg, et Brad Price. 2023. *car: Companion to Applied Regression*. <https://CRAN.R-project.org/package=car>.
- Horst, Allison, Alison Hill, et Kristen Gorman. 2022. *palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. <https://CRAN.R-project.org/package=palmerpenguins>.
- Liberg, Olof, Henrik Andrén, Hans-Christian Pedersen, Håkan Sand, Sejberg, Petter Wabakken, Mikael Åkesson, et Staffan Bensch. 2005. « Severe Inbreeding Depression in a Wild Wolf *Canis Lupus* Population ». *Biology Letters* 1 (1): 17-20. <https://doi.org/10.1098/rsbl.2004.0266>.
- Molofsky, Jane, et Jean-Baptiste Ferdy. 2005. « Extinction dynamics in experimental metapopulations ». *Proceedings of the National Academy of Sciences* 102 (10): 3726-31. <https://doi.org/10.1073/pnas.0404576102>.
- Müller, Martina S., Elaine T. Porter, Jacquelyn K. Grace, Jill A. Awkerman, Kevin T. Birchler, Alex R. Gunderson, Eric G. Schneider, Mark A. Westbrock, et David J. Anderson. 2011. « Maltreated Nestlings Exhibit Correlated Maltreatment as Adults: Evidence of a “Cycle of Violence” in Nazca Boobies (*Sula Granti*) ». *The Auk* 128 (4): 615-19. <https://doi.org/10.1525/auk.2011.11008>.

- Robinson, David, Alex Hayes, et Simon Couch. 2023. *broom: Convert Statistical Objects into Tidy Tibbles*. <https://CRAN.R-project.org/package=broom>.
- Signorell, Andri. 2023. *DescTools: Tools for Descriptive Statistics*. <https://CRAN.R-project.org/package=DescTools>.
- Tilman, David, Peter B. Reich, et Johannes M. H. Knops. 2006. « Biodiversity and Ecosystem Stability in a Decade-Long Grassland Experiment ». *Nature* 441 (7093): 629. <https://doi.org/10.1038/nature04742>.
- Waring, Elin, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, et Shannon Ellis. 2022. *skimr: Compact and Flexible Summaries of Data*. <https://CRAN.R-project.org/package=skimr>.
- Whitlock, Michael, et Dolph Schluter. 2015. *The Analysis of Biological Data*. Second edition. Greenwood Village, Colorado: Roberts and Company Publishers.
- Wickham, Hadley. 2023. *tidyverse: Easily Install and Load the Tidyverse*. <https://CRAN.R-project.org/package=tidyverse>.
- Wickham, Hadley, et Jennifer Bryan. 2023. *readxl: Read Excel Files*. <https://CRAN.R-project.org/package=readxl>.
- Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, et Dewey Dunnington. 2023. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, et Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, Jim Hester, et Jennifer Bryan. 2023. *readr: Read Rectangular Text Data*. <https://CRAN.R-project.org/package=readr>.
- Wiseman, Richard, et Peter Lamont. 1996. « Unravelling the Indian Rope-Trick ». *Nature* 383 (6597): 212. <https://doi.org/10.1038/383212a0>.
- Wright, Kenneth P., et Charles A. Czeisler. 2002. « Absence of Circadian Phase Resetting in Response to Bright Light Behind the Knees ». *Science* 297 (5581): 571-71. <https://doi.org/10.1126/science.1071697>.