

TP de Biométrie Semestre 5

Benoît Simon-Bouhet

jeudi 6 juillet 2023

Table des matières

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 4*' du semestre 5 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, etc.)
- 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
- Identifier des cohortes dans une population et en étudier les caractéristiques et l'évolution temporelle

- Simuler le comportement de populations théoriques simples suivant des modèles démographiques précis (mortalité exponentielle, croissance exponentielle, croissance logistique, système prédateur-proies de Lotka et Volterra, et systèmes de compétition à 2 ou 3 espèces...)
- Simuler, par chaînes de Markov, les successions écologiques dans un écosystème théorique

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 S5, je suppose que vous possédez 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.

 Si ces pré-requis ne sont pas maîtrisés

- mettez-vous à niveau de toute urgence en lisant attentivement [le livre en ligne de Biométrie du semestre 3](#)
- mettez-vous en binôme avec un · e collègue qui a suivi l'EC Immersion R et RStudio en début de semestre. Ça ne vous dispensera pas de lire le livre en ligne de Biométrie S3, mais ça vous fera certainement gagner pas mal de temps.

Organisation

Volume de travail

Les travaux pratiques et TEA de biométrie auront lieu entre le 17 octobre et le 02 décembre 2022 :

- Semaine 42 (du 17 au 21 octobre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 43 (du 24 au 28 octobre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 45 (du 07 au 10 novembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 46 (du 14 au 18 novembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 47 (du 21 au 25 novembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 48 (du 28 novembre au 02 décembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30

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

Au total, chaque groupe aura donc 6 séances de TP et 6 séances de TEA, soit un total de 18 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 vingtaine d'heures de travail personnel en plus des séances prévues dans votre emploi du temps. 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 allez en effet avoir des enseignements qui reposent sur l'utilisation de ces logiciels jusqu'à la fin du S6 (y compris pendant vos 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 l'EC immersion 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. 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, que vous n'avez pas de problème avec les exercices de ce livre en ligne ni avec les quizz Moodle, 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.

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, votre venue sera totalement inutile puisque vous n'aurez pas de question à poser et que vous passerez votre séance à lire et suivre ce livre en ligne, choses que vous pouvez très bien faire chez vous. Vous perdrez donc votre temps, celui de vos collègues, et le mien. De même, si vous attendez la 4e semaine pour vous y mettre, vous irez droit dans le mur. Je le répète, outre les heures de TP/TEA prévus dans vos emplois du temps, vous devez prévoir au moins 20 heures de travail personnel supplémentaire.

Je vous laisse donc une grande liberté d'organisation. À vous d'en tirer le maximum et de faire preuve du sérieux nécessaire. Le rythme auquel vous devriez avancer est présenté dans la partie suivante intitulée "Progression conseillée".

Utilisation de Slack

Outre les séances de permanence non-obligatoires, nous échangerons aussi sur [l'application Slack](#), qui fonctionne un peu comme un "twitter privé". Slack facilite la communication des équipes et permet de travailler ensemble. 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).

Vous verrez que 3 "chaînes" 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 le document de prise en main de R et RStudio (lors de l'immersion ou lors d'une remise à niveau en autonomie), vous savez que pour apprendre à utiliser ces logiciels, 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 (42) est consacrée l'exploration statistique des jeux de données. Avant votre seconde séance de TP, vous devriez avoir compris comment calculer et interpréter des résumés statistiques de vos jeux de données. Vous devriez en particulier être capable de calculer des estimateurs de position (moyennes, médianes, quartiles...) et de dispersion (variances, écart-types, intervalles inter-quartiles...) sur des variables numériques, et ce, pour plusieurs modalités d'une variable catégorielle ou pour chaque combinaison de modalités de plusieurs variables catégorielles (par exemple, quelles sont les moyennes et variances des longueurs de becs pour chaque espèce de manchots et chaque sexe). Vous devrez être capables de distinguer la notion de dispersion de celle de précision, et vous devrez être capables de calculer l'erreur standard de la moyenne (ou erreur type). Vous devrez en outre être capables de produire des graphiques sur lesquels apparaissent des barres d'incertitude (erreurs standards ou intervalles de confiance).
2. La deuxième semaine (43) est consacrée aux tests statistiques. Avant votre troisième séance de TP, vous devriez être capable de formuler des hypothèses nulles et alternatives pertinentes, et vous devriez connaître le concept de p -value. Vous devriez en outre être capables, avec des données de comptages, de réaliser des tests de comparaison de proportions, et d'en interpréter correctement les résultats.
3. La troisième semaine (45) est également consacrée aux tests d'hypothèses. Avant votre quatrième séance de TP, vous devriez être capable de comparer la moyenne d'une population à une valeur théorique, et de comparer la moyenne de 2 populations, dans le cas où vous disposez d'échantillons appariés. Dans les deux cas, vous devrez être capable de vérifier les conditions d'application des tests paramétriques, et de choisir des tests non-paramétriques équivalents si les conditions d'application ne sont pas vérifiées.
4. La quatrième semaine (46) est consacrée aux derniers tests de comparaison de moyennes. Avant votre cinquième séance de TP, vous devrez donc être capable de comparer la moyenne de deux populations lorsque

les échantillons sont indépendants. Comme pour la semaine précédente, vous devrez être capable de vérifier les conditions d'application du test paramétrique, et de réaliser les tests non paramétrique équivalent le cas échéant. Enfin, vous devrez aussi être en mesure de spécifier les hypothèses alternatives unilatérales ou bilatérales pertinentes selon la question scientifique posée. Pour chaque semaine consacrée aux tests, vous devrez aussi toujours penser à examiner les données graphiquement, et par le biais des statistiques descriptives décrites lors de la première semaine

5. La cinquième semaine (47) est consacrée à la mise en pratique des notions vues dans le cours magistral de Population Dynamics (EC "Fonctionnement des Écosystèmes). Nous aborderons ici les analyses de cohorte. Avant votre dernière séance de TP, vous devriez être en mesure de réaliser l'analyse de cohorte d'une population étudiée pendant plusieurs années afin de produire les courbes de croissance, de survie et d'Allen d'une cohorte d'intérêt. Vous devrez en particulier importer et mettre en forme des données issues d'un suivi de terrain, produire les structures démographiques instantanées à chaque date d'échantillonnage, faire les décompositions polymodales afin de récupérer les informations utiles au sujet de la cohorte dont on souhaite assurer le suivi.
6. La sixième semaine (48) est consacrée à la mise en pratique des notions vues dans le cours magistral de Population Dynamics (EC "Fonctionnement des Écosystèmes). Nous aborderons ici l'étude des systèmes dynamiques. Vous devrez coder différents modèles d'évolution d'une populations (mortalité exponentielle, croissance exponentielle, croissance logistique) ou de plusieurs populations ou espèces (modèle prédateurs-proies, modèle de compétition). Ces modèles généreront des données que vous devrez représenter graphiquement. Vous devrez enfin modifier la valeur de certains paramètres de ces modèles afin de comprendre leur influence sur le comportement des systèmes dynamiques étudiés.

Évaluation(s)

L'évaluation de la partie "Biométrie" de l'EC "Outils pour l'étude et la compréhension du vivant" aura lieu dans le cadre du travail de stratégie d'échantillonnage que vous mettez en œuvre avec Pierrick Bocher. Le compte-rendu de stratégie d'échantillonnage servira donc à évaluer 3 choses :

- les grands principes de stratégie d'échantillonnage abordés par Pierrick
- la mise en œuvre de méthodes statistiques adaptées pour répondre aux questions scientifiques posées, telles que nous les traitons en Biométrie
- la maîtrise du 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). Pour ce dernier volet, vous devrez rendre, en plus de votre compte-rendu, votre script d'analyse. C'est ce script qui me permettra d'évaluer votre niveau de compétence et de maîtrise de l'outil, tant sur la forme du script (lisibilité, structure, reproductibilité, etc.) que sur le fond (pertinence des analyses réalisées).

Pour vous aider à comprendre ce qui est attendu, 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"

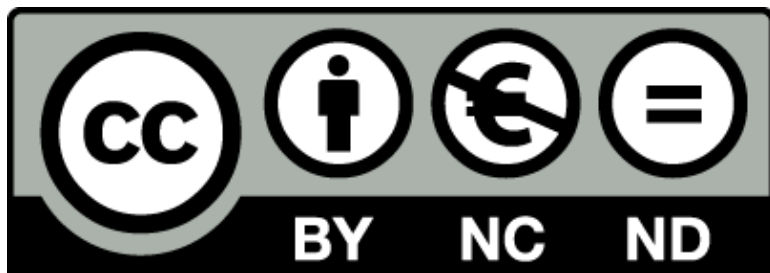
Grille d'évaluation

	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 inadaptes	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	

Pour ce qui est du fond (pertinence des analyses statistiques réalisées et de leurs interprétations), une autre grille critériée sera fournie ici avant la fin du semestre.

Licence

Ce livre est sous licence Creative Commons ([CC BY-NC-ND 4.0](https://creativecommons.org/licenses/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 la moyenne d'une population à une valeur théorique

1.1 Pré-requis

Pour travailler dans de bonnes conditions, créez un nouveau dossier sur votre ordinateur, créez un **Rproject** et un script dans ce dossier, et travaillez systématiquement **dans votre script**, et surtout pas directement dans la console. Consultez [le livre en ligne du semestre 3](#) si vous ne savez plus comment faire.

Dans ce chapitre, vous aurez besoin d'utiliser des packages spécifiques et d'importer des données depuis des fichiers externes téléchargeables directement depuis ce document. Les packages dont vous aurez besoin ici et que vous devez donc charger en mémoire, sont :

- 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)
```

⚠ Important

Même si vous avez déjà installé le `tidyverse` ou `dplyr` au semestre précédent, ré-installez `dplyr` avec `install.packages("dplyr")`. Ce package a en effet été mis à jour tout récemment, et nous aurons besoin de sa toute dernière version (v1.1.0). Chargez-le ensuite en mémoire avec `library(dplyr)`.

! Attention

Pensez à installer tous les packages listés ci-dessous avant de les charger en mémoire si vous ne l'avez pas déjà fait. Si vous ne savez plus comment faire, consultez d'urgence [la section dédiée aux packages dans le livre en ligne de Biométrie du semestre 3](#).

Vous aurez également besoin des jeux de données suivants :

- [Temperature.csv](#)
- [Temperature2.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

On s'intéresse ici à la température corporelle des adultes en bonne santé. On souhaite examiner la croyance populaire qui veut que cette température vaut en moyenne 37°C. Pour le vérifier, on dispose d'un échantillon de 25 adultes en bonne santé choisis au hasard parmi la population américaine et dont on a mesuré la température. Comme pour toute étude statistique, les étapes que nous allons devoir suivre sont les suivantes (dans l'ordre) :

1. Importer les données dans **RStudio**, les examiner et éventuellement les (re)mettre en forme si besoin.
2. Faire une première exploration des données, grâce au calcul d'indices de statistiques descriptives d'une part, et de représentations graphiques d'autre part.
3. Réaliser un test d'hypothèses en respectant la procédure adéquate (en particulier, la vérification des conditions d'application).

C'est donc ce que nous allons faire dans les sections suivantes.

! À retenir !

Avant de se lancer dans les tests d'hypothèses, il est **toujours indispensable** d'examiner les données dont on dispose à l'aide, d'une part de statistiques descriptives numériques, et d'autre part, de graphiques exploratoires.

Nous avons vu au cours des semestres précédents quels indices statistiques il peut être utile de calculer ([dans le livre en ligne du semestre 4](#)) et quelles représentations graphiques il peut être utile de réaliser ([dans le livre en ligne du semestre 3](#)) afin de pouvoir se lancer dans des tests d'hypothèses sans risquer de grossières erreurs. N'hésitez pas à cliquer sur ces liens pour vous rafraîchir la mémoire !

1.3 Importation et mise en forme des données

Nous allons travailler ici sur les données contenues dans le fichier [Temperature.csv](#). Téléchargez ces données dans votre répertoire de travail (attention : ne les ouvrez pas avec Excel !), puis importez les données dans **RStudio** grâce à l'assistant d'importation. Si vous ne savez plus comment faire, consultez [la section dédiée à l'importation des données dans le livre en ligne de Biométrie du semestre 3](#)

Vous stockerez les données dans un objet que vous nommerez **Temperature**. Après l'importation, tapez son nom dans

la console de RStudio et vérifiez que vous obtenez bien exactement ce résultat :

```
Temperature

# A tibble: 25 x 2
  individual temperature
      <dbl>         <dbl>
1         1         98.4
2         2         98.6
3         3         97.8
4         4         98.8
5         5         97.9
6         6          99
7         7         98.2
8         8         98.8
9         9         98.8
10        10          99
# i 15 more rows
```

La première chose à faire quand on travaille avec des données inconnues, c'est d'examiner les données brutes. Ici, les données sont importées au format `tibble`, donc seules les premières lignes sont visibles. Pour visualiser l'ensemble du tableau, utilisez la fonction `View()` (avec un `V` majuscule) ou, si vous avez mis en mémoire le `tidyverse`, la fonction `view()` (sans majuscule) :

```
View(Temperature)
```

Cette commande ouvre un nouvel onglet présentant les données dans un tableur simplifié, en lecture seule. On constate ici 2 choses que nous allons modifier :

1. la première colonne, intitulée `individual`, n'est pas véritablement une variable. Cette colonne ne contient qu'un identifiant sans intérêt pour notre étude et est en fait identique au numéro de ligne. Nous allons donc supprimer cette colonne.
2. les températures sont exprimées en degrés Fahrenheit, ce qui rend leur lecture difficile pour nous qui sommes

habitué à utiliser le système métrique et les degrés Celsius. Nous allons donc convertir les températures en degrés Celsius grâce à la formule suivante :

$$C = \frac{F - 32}{1.8}$$

```
Temp_clean <- Temperature %>%  
  select(-individual) %>% # Suppression de la colonne `individual`  
  mutate(                 # Transformation des températures en °Celsius  
    temperature = (temperature - 32) / 1.8  
  )  
  
Temp_clean
```

```
# A tibble: 25 x 1  
  temperature  
    <dbl>  
1      36.9  
2      37  
3      36.6  
4      37.1  
5      36.6  
6      37.2  
7      36.8  
8      37.1  
9      37.1  
10     37.2  
# i 15 more rows
```

Il nous est maintenant possible d'examiner à nouveau les données avec la fonction `View()`. Avec des valeurs de températures comprises entre 36.3°C et 37.8°C, il n'y a visiblement pas de données aberrantes.

Examiner les données brutes est donc la première chose que vous devriez prendre l'habitude de faire, et ce de façon systématique, car cela permet de repérer :

- La nature des variables présentes.
- Les variables inutiles qui pourront être supprimées ou négligées.

- Les unités des variables utiles, afin de pouvoir les convertir si nécessaire.
- Les valeurs manquantes, atypiques ou aberrantes qui demanderont toujours une attention particulière.

Maintenant que l'examen préliminaire des données est réalisé, on peut passer au calcul des statistiques descriptives.

1.4 Exploration statistique des données

1.4.1 Position et dispersion

On s'intéresse ici au calcul de grandeurs statistiques nous apportant des renseignements sur la **position** et la **dispersion** des valeurs de l'échantillon. Les questions auxquelles on tente de répondre à ce stade sont les suivantes :

- Quelle est la tendance centrale (moyenne ou médiane) ?
- Quelle est la dispersion des valeurs autour de la tendance centrale (écart-type, variance, intervalle interquartile...) ?

Pour répondre à ces questions, on peut faire appel à de multiples fonctions déjà présentées [dans le livre en ligne du semestre 4](#). Par exemple la fonction `summarise()`, en conjonction avec les fonctions `mean()`, `median()`, `sd()`, `var()`, `min()`, `max()` ou `quantile()`, ou les fonctions `summary()` ou `skim()` (du package `skimr`).

Je prends ici un exemple simple, mais n'hésitez pas à expérimenter avec les méthodes décrites dans le livre en ligne du semestre 4.

```
summary(Temp_clean)
```

```
temperature
Min.      :36.33
1st Qu.   :36.67
Median    :37.00
Mean      :36.96
3rd Qu.   :37.22
Max.      :37.78
```

On constate ici que la moyenne et la médiane sont très proches. La distribution des températures doit donc être à peu près symétrique, avec à peu près autant de valeurs au-dessus que de valeurs en dessous de la moyenne. Les premier et troisième quartiles sont à peu près aussi éloignés de la médiane l'un que l'autre, ce qui confirme l'apparente symétrie du jeu de données de part et d'autre de la tendance centrale.

La moyenne observée dans l'échantillon vaut 36.96°C, ce qui est très proche de la moyenne théorique de 37°C.

Une autre fonction utile est la fonction `IQR()`, qui renvoie l'étendue de l'intervalle interquartile (la valeur du troisième quartile moins la valeur de premier quartile) :

```
Temp_clean %>%
  summarise(IQ_range = IQR(temperature))

# A tibble: 1 x 1
  IQ_range
    <dbl>
1      0.556
```

On constate ici que l'intervalle interquartile a une largeur de 0.56°C. Cela signifie que les 50% des températures les plus centrales de l'échantillon sont situées dans un intervalle d'environ un demi-degré Celsius autour de la médiane.

Enfin, pour obtenir des informations complémentaires, on peut utiliser la fonction `skim()` du package `skimr` :

```
skim(Temp_clean)

-- Data Summary -----

```

	Values
Name	Temp_clean
Number of rows	25
Number of columns	1

```
-----
Column type frequency:
  numeric      1
```

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

```
-- Variable type: numeric -----  
  skim_variable n_missing complete_rate mean    sd   p0  p25 p50  p75 p100 hist  
1 temperature           0             1 37.0 0.377 36.3 36.7  37 37.2 37.8
```

Tout comme `summary()`, la fonction `skim()` renvoie les valeurs minimales et maximales, les premiers et troisièmes quartiles ainsi que la moyenne et la médiane. Elle nous indique en outre la valeur de l'écart-type de l'échantillon, ainsi que le nombre d'observations et le nombre de données manquantes. Enfin, elle fournit un histogramme très simplifié et sans échelle. Cet histogramme nous permet de nous faire une première idée de la distribution des données et est particulièrement utile pour comparer rapidement un grand nombre de distributions quand il y a plusieurs catégories dans les données (ce qui n'est pas le cas ici).

Outre ces 3 fonctions (`summary()`, `IQR()`, et `skim()`), il est bien sûr possible de calculer toutes ces valeurs manuellement si besoin :

- `mean()` permet de calculer la moyenne.
- `median()` permet de calculer la médiane.
- `min()` et `max()` permettent de calculer les valeurs minimales et maximales respectivement.
- `quantile()` permet de calculer les quartiles.
- `sd()` permet de calculer l'écart-type.
- `var()` permet de calculer la variance.
- `n()` permet de compter le nombre d'observations.

Toutes ces fonctions prennent seulement un vecteur en guise d'argument. Il faut donc procéder comme avec `IQR()` pour les utiliser, en les intégrant à l'intérieur de la fonction `summarise()`. Par exemple, pour calculer la variance, on peut taper :

```
Temp_clean %>%  
  summarise(variance = var(temperature))
```

```
# A tibble: 1 x 1  
  variance
```

```
      <dbl>
1      0.142
```

ou :

```
Temp_clean %>%
  pull(temperature) %>%
  var()
```

```
[1] 0.1417901
```

ou encore :

```
var(Temp_clean$temperature)
```

```
[1] 0.1417901
```

À vous d'utiliser la syntaxe qui vous semble la plus simple.

1.4.2 Incertitude

Outre les informations de position et de dispersion, nous avons vu [au semestre 4](#) qu'il était également important d'avoir une idée de l'**incertitude** associée aux estimations de tendance centrale (erreur standard ou intervalle de confiance de la moyenne ou médiane). Ici, nous allons donc calculer l'intervalle de confiance à 95% de la moyenne. Si vous ne savez plus comment faire, ou que vous ne comprenez pas le code ci-dessous, consultez [le livre en ligne du semestre 4](#) :

```
Temp_clean %>%
  reframe(mean_cl_normal(temperature))
```

```
# A tibble: 1 x 3
      y ymin ymax
  <dbl> <dbl> <dbl>
1  37.0  36.8  37.1
```

On constate ici que les bornes inférieure (36.8°C) et supérieure (37.1°C) de l'intervalle de confiance à 95% de la moyenne sont proches de la valeur de moyenne de l'échantillon. Dans la population générale, la moyenne de la température corporelle chez les adultes en bonne santé a de bonnes chances de se trouver quelque part entre 36.8°C et 37.1°C. Autrement dit, si la température corporelle des adultes en bonne santé n'est pas exactement de 37°C, l'écart à cette valeur théorique ne doit pas être très important.

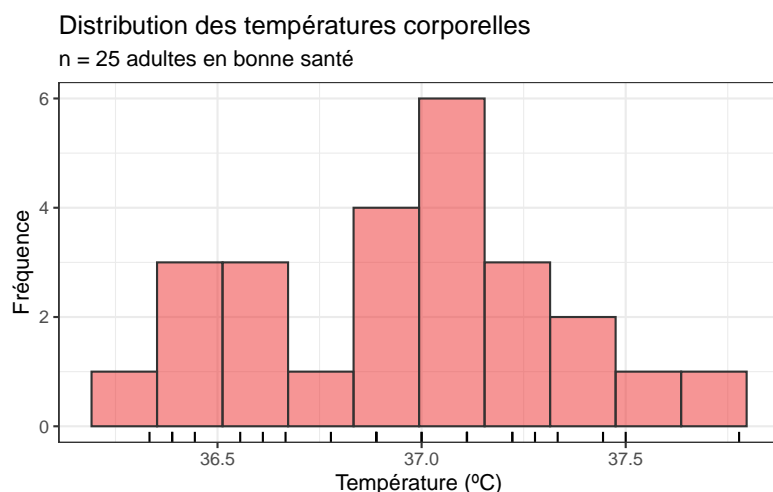
1.5 Exploration graphique des données

Ici, puisque nous ne disposons que d'une unique variable numérique et que nous n'avons donc qu'un unique groupe, les représentations graphiques que nous allons réaliser doivent nous permettre d'examiner la **distribution des données**. Pour cela, nous pouvons réaliser soit un histogramme, soit un diagramme de densité.

1.5.1 Histogramme

Voilà comment produire un histogramme de qualité pour les données de températures :

```
Temp_clean %>%
  ggplot(aes(x = temperature)) +
  geom_histogram(bins = 10, fill = "firebrick2", color = "grey20",
                 alpha = 0.5) +
  geom_rug() +
  labs(x = "Température (°C)",
       y = "Fréquence",
       title = "Distribution des températures corporelles",
       subtitle = "n = 25 adultes en bonne santé")
```

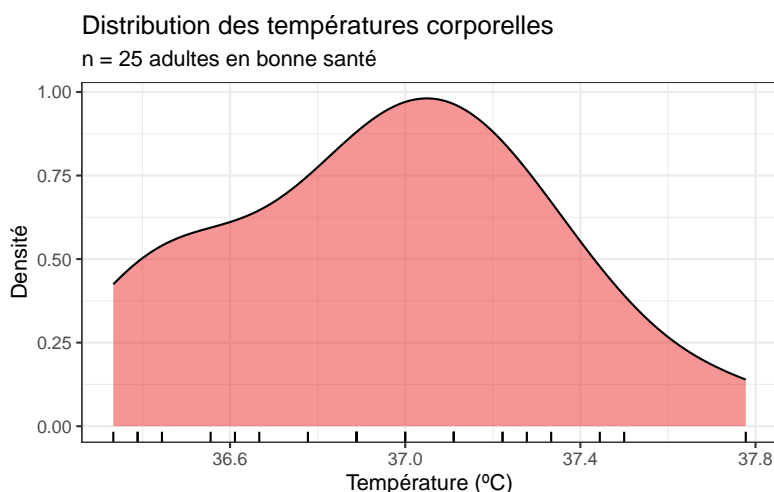
Si vous ne vous rappelez-plus ce qu'est un histogramme ou comment le faire, ou la signification de l'argument `bins`, relisez [la section consacrée aux histogrammes](#) du livre en ligne de Biométrie du semestre 3. Notez que j'ai ajouté une couleur de remplissage et de la transparence pour rendre le graphique plus facile à lire. J'ai également spécifié des titres pour les axes (en précisant l'unité de la variable numérique dont on représente la distribution) ainsi que le titre (et sous-titre) du graphique, qui précise ce qu'on a sous les yeux et la taille de l'échantillon. Il n'est pas toujours nécessaire de spécifier le titre (et le sous-titre) de cette façon : lorsque vous intégrez des graphiques dans un compte-rendu ou un rapport, le titre est en général précisé sous la figure, au début d'une légende qui la décrit. Enfin, j'ai ajouté `geom_rug()` pour faire apparaître sous le graphique, le long de l'axe des x , la position des données observées. Cela permet de visualiser les données brutes, et peut donc permettre de mieux comprendre pourquoi un histogramme présente telle ou telle forme.

Ici, la forme de ce l'histogramme est assez proche de celle présentée plus tôt par l'histogramme très simplifié produit par la fonction `skim()`. Cet histogramme nous apprend qu'en dehors d'un "trou" autour de la température 36.75°C, la distribution des données est proche d'une courbe en cloche. Il y a fort à parier qu'un test de normalité conclurait à la normalité des données de cet échantillon. C'est ce que nous verrons dans la Section ??.

1.5.2 Diagramme de densité

Une autre façon de visualiser la distribution d'une variable numérique est de produire un graphique de densité. Il a l'avantage d'éviter à l'utilisateur d'avoir à choisir une valeur pour l'argument `bin` de la fonction `geom_histogram()`, mais il a l'inconvénient de présenter une échelle plus difficile à comprendre pour l'axe des ordonnées :

```
Temp_clean %>%  
  ggplot(aes(x = temperature)) +  
  geom_density(fill = "firebrick2", alpha = 0.5) +  
  geom_rug() +  
  labs(x = "Température (°C)",  
       y = "Densité",  
       title = "Distribution des températures corporelles",  
       subtitle = "n = 25 adultes en bonne santé")
```



Les informations apportées par ce graphique sont cohérentes avec celle de l'histogramme :

- les températures les plus fréquemment observées dans notre échantillon de 25 adultes en bonne santé se situent légèrement au dessus de 37°C. Il s'agit d'une information concernant la **position** des données (c'est-à-dire où se trouve le pic de la distribution sur l'axe des x)
- les températures observées ont une distribution qui ressemble à peu près à une courbe en cloche, avec des

valeurs comprises entre 36.4°C et 37.8°C environ. La symétrie de part et d'autre du pic n'est pas parfaite, mais elle reste bonne. Il s'agit d'informations concernant la forme de la distribution et la **dispersion** des données.

i Bilan des analyses préliminaires

Suite à l'exploration statistique et graphique des données de températures, voilà ce qu'on retient :

1. Il n'y a visiblement pas de données aberrantes.
2. La distribution des données semble suivre à peu près la loi Normale.
3. La médiane et la moyenne sont très proches de 37°C. Un test devrait donc arriver à la conclusion que la température corporelle des adultes n'est pas significativement différente de 37°C.
4. La largeur de l'intervalle de confiance à 95% semble faible, ce qui indique une incertitude relativement faible. Si la température réelle des adultes en bonne santé n'est pas exactement de 37°C, elle ne devrait pas en être très éloignée (quelques dixièmes de degrés Celsius au plus).

1.6 Le test paramétrique

Le test permettant de comparer la moyenne μ d'une population à une valeur théorique, fixée par l'utilisateur, est le **test de Student à un échantillon**. Il permet de répondre à la question suivante :

Les données observés dans l'échantillon dont je dispose sont-elles compatibles avec l'hypothèse que la moyenne μ de la population dont est issu mon échantillon vaut **XXX** ?

avec **XXX**, une valeur d'intérêt spécifiée par l'utilisateur. Il s'agit d'un test paramétrique très puissant. Comme tous les tests paramétriques, certaines conditions d'application doivent être vérifiées avant de pouvoir l'appliquer.

! Important

Comme pour tous les tests statistiques que nous allons réaliser lors de ces séances de TP et TEA, nous devons commencer par **spécifier les hypothèses** nulles et alternatives de chaque test, ainsi que la **valeur du seuil** α que nous allons utiliser. À moins d'avoir une bonne raison de faire autrement, on utilise presque toujours le seuil $\alpha = 0.05$ dans le domaine des sciences du vivant. C'est donc ce seuil que nous utiliserons dans ce livre en ligne.

1.6.1 Conditions d'application

Les conditions d'application du test de Student à un échantillon sont les suivantes :

1. Les données de l'échantillon sont issues d'un **échantillonnage aléatoire** au sein de la population générale. Cette condition est partagée par toutes les méthodes que nous verrons dans ces TP. En l'absence d'informations sur la façon dont l'échantillonnage a été réalisé, on considère que cette condition est remplie. Il n'y a pas de moyen statistique de le vérifier, cela fait uniquement référence à la stratégie d'échantillonnage déployée et à la rigueur de la procédure mise en œuvre lors de l'acquisition des données.
2. La variable étudiée doit suivre une **distribution Normale** dans la population générale. Nous allons vérifier cette condition d'application avec un **test de normalité de Shapiro-Wilk**.

Pour un test de normalité, les hypothèses seront toujours les suivantes :

- H_0 : la variable étudiée suit une distribution Normale dans la population générale.
- H_1 : la variable étudiée ne suit pas une distribution Normale dans la population générale.

Le test de Shapiro-Wilk se réalise de la façon suivante :

```
shapiro.test(Temp_clean$temperature)
```

ou

```
Temp_clean %>%  
  pull(temperature) %>%  
  shapiro.test()
```

Shapiro-Wilk normality test

```
data: .  
W = 0.97216, p-value = 0.7001
```

la fonction `pull()` permet d'extraire une colonne (ici `temperature`) d'un tibble (ici `Temp_clean`) et de la transformer en vecteur.

W est la statistique du test. Elle permet à RStudio de calculer la p -value. Ici, $p > \alpha$. On ne peut donc pas rejeter l'hypothèse nulle de normalité : on ne peut pas exclure que dans la population générale, la température suive bel et bien une distribution Normale. Les conditions d'application du test de Student sont bien vérifiées.

⚠ Tests et décision : rappel de cours


À l'issue d'un tests statistique, la décision finale est toujours prise par rapport à l'hypothèse nulle (H_0) :

- Si la p -value du test est supérieure ou égale à α , on dit qu'on ne peut pas rejeter l'hypothèse nulle H_0 . Attention, on ne dit jamais que " H_0 est vraie", car il est impossible de le vérifier avec une certitude absolue. Toutefois, les données observées (celles de notre échantillon), sont compatibles avec l'hypothèse nulle que nous avons formulée, jusqu'à preuve du contraire.
- Si la p -value du test est inférieure à α , on dit qu'on rejette l'hypothèse nulle au seuil α . Autrement dit, les données observées ne sont pas compatibles avec l'hypothèse nulle. On accepte alors l'hypothèse alternative (H_A).

L'hypothèse nulle est toujours l'hypothèse la moins “intéressante”, celle pour laquelle “il ne se passe rien de notable” (par exemple : “les données suivent la distribution Normale”, ou “les moyennes sont égales”).

1.6.2 Signification de la p -value

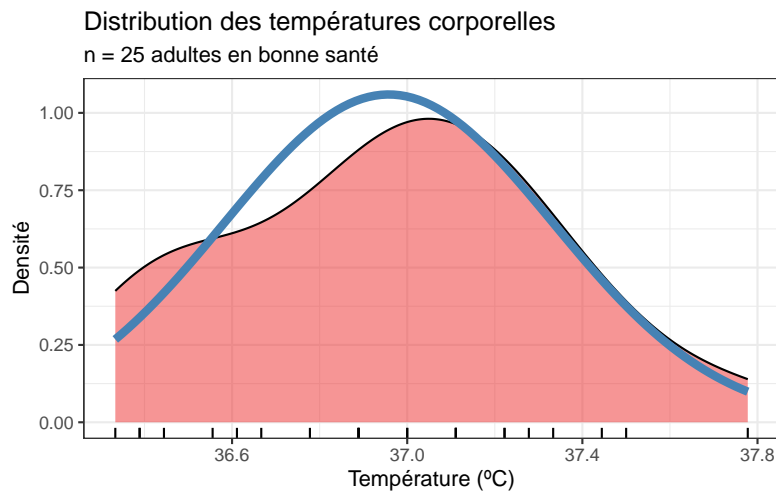
La p -value est une grandeur centrale en statistiques et elle est souvent mal comprise et donc mal interprétée. Je prends donc le temps ici d'expliquer ce qu'est la p -value et comment il faut la comprendre.

 Définition : la p -value

La p -value d'un test statistique, c'est la probabilité, si H_0 est vraie, d'obtenir un effet au moins aussi extrême que celui qu'on a observé dans l'échantillon, sous le seul effet du hasard.

Ici, la p -value de notre test de Normalité de Shapiro-Wilk vaut 0.7101. Cela signifie que si les données suivent réellement la loi Normale dans la population générale (donc si H_0 est vraie), l'écart à la Normalité que nous avons observé (ou un écart encore plus important), peut être observé dans 70.1% des cas. Autrement dit, si on prélève un grand nombre d'échantillons de 25 adultes dans la population générale et qu'on regarde à quoi ressemble la distribution des températures dans chacun de ces échantillons, pour 70.1% d'entre eux, la distribution obtenue sera au moins aussi éloignée de la distribution Normale que celle que nous avons observée ici.

Dans notre cas, l'écart entre la loi Normale et les données de notre échantillon peut être visualisé de la façon suivante :

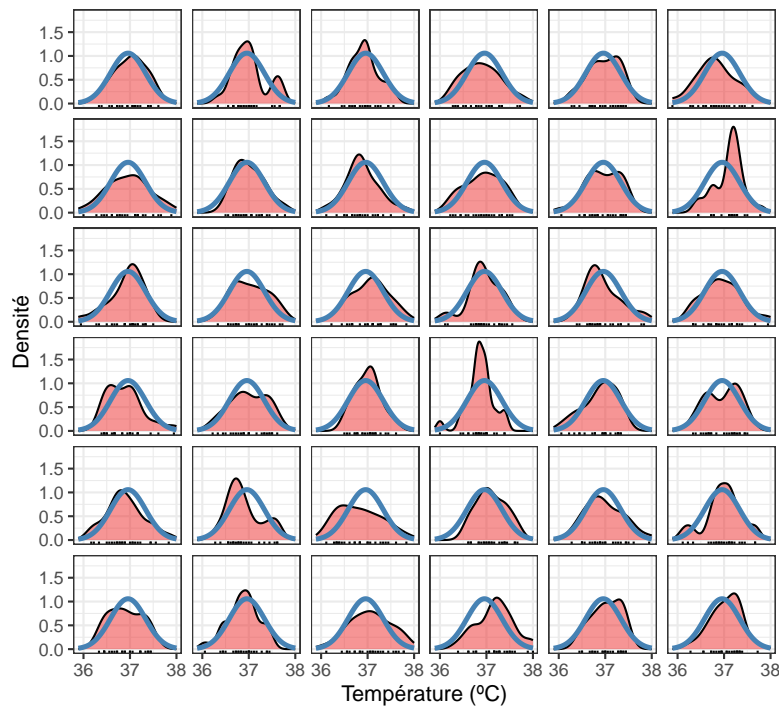


La courbe de densité des données observées est en rouge, et la distribution Normale théorique correspond à la courbe en bleu. Il y a donc un écart entre la courbe en cloche parfaite de la loi Normale et les données observées. La p -value du test de Shapiro-Wilk nous dit que si la température des adultes en bonne santé suit réellement la loi Normale dans la population générale, alors, l'écart que nous avons observé, ou un écart encore plus important, peut être observé simplement par hasard dans 70.1% des cas. Autrement dit, c'est très probable, et on peut donc considérer que l'écart à la loi Normale que nous avons observé est le fruit du hasard et que notre variable suit donc bien la Loi Normale.

Pour bien comprendre cette notion importante, je simule ci-dessous 36 échantillons de 25 adultes dont les températures suivent parfaitement la loi Normale dans la population générale. Je me place donc dans la situation où je sais que H_0 est vraie, pour illustrer la notion de *fluctuation d'échantillonnage*. En raison du seul hasard de l'échantillonnage, et alors même que les échantillons que je génère sont issus d'une population qui suit parfaitement la Normale, la distribution dans chaque échantillon s'écarte parfois fortement de la courbe en cloche théorique :

Distribution des températures corporelles

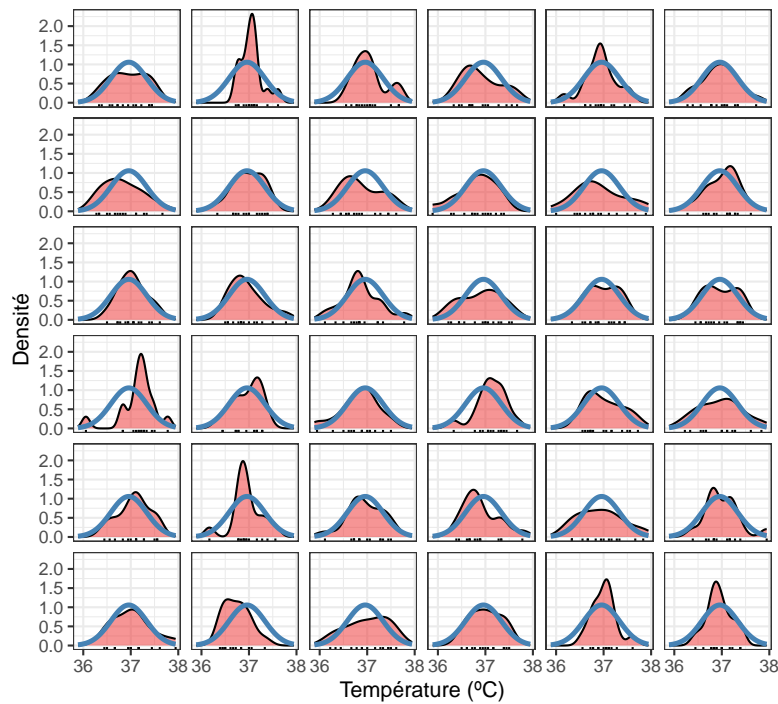
36 échantillons de $n = 25$ adultes en bonne santé



On voit bien ici que certains échantillons s'écartent fortement de la distribution théorique alors même que tous les échantillons sont issus d'une population Normale. Et plus l'échantillon sera de taille réduite, plus les écarts à la courbe en cloche parfaite seront grands. La preuve ci-dessous avec des échantillons de $n = 15$ adultes au lieu de 25 :

Distribution des températures corporelles

36 échantillons de $n = 15$ adultes en bonne santé



Au final, la p -value de 0.701 de notre test de Shapiro-Wilk nous indique que l'hypothèse de la Normalité n'est pas incompatible avec les données que nous avons observées.

Imaginons qu'à l'inverse, nous ayons obtenu une p -value très faible, égale à 0.01 par exemple (donc inférieure à notre seuil α de 0.05). Nous aurions alors rejeté l'hypothèse nulle. En effet, obtenir une p -value de 0.01, signifie que si H_0 est vraie, obtenir un écart à la courbe en cloche théorique aussi important que celui que nous observons est très peu probable (une chance sur 100). Puisqu'il est très improbable d'observer un tel écart si H_0 est vraie, on en conclut que H_0 n'est pas vraie : les données sont incompatibles avec l'hypothèse nulle et on la rejette donc logiquement.

Cette logique sera valable pour tous les autres tests statistiques que nous aborderons dans cet ouvrage. Pour un test de Normalité, on regarde l'écart entre la distribution Normale et les données observées. Pour un test de comparaison de moyennes, on regarde l'écart entre la moyenne théorique et la moyenne observée, ou entre les 2 moyennes qu'on essaie de comparer. Mais la philosophie reste la même.

1.6.3 Réalisation du test de Student et interprétation

Puisque les conditions d'application du test de Student à un échantillon sont vérifiées, nous avons le droit de faire ce test, et nous devons donc maintenant spécifier les hypothèses nulles et alternatives que nous allons utiliser pour le réaliser :

- H_0 : dans la population générale, la température corporelle moyenne des adultes en bonne santé vaut 37°C ($\mu = 37$).
- H_1 : dans la population générale, la température corporelle moyenne des adultes en bonne santé est différente de 37°C ($\mu \neq 37$).

Hypothèses et paramètres

Notez que les hypothèses des tests statistiques concernent toujours la valeur d'un paramètre de la population générale, et non la valeur des estimateurs calculés dans un échantillon.

On réalise ensuite le test de la façon suivante :

```
t.test(Temp_clean$temperature, mu = 37)
```

ou

```
t.test(temperature ~ 1, mu = 37, data = Temp_clean)
```

ou encore,

```
Temp_clean %>%  
  pull(temperature) %>%  
  t.test(mu = 37)
```

One Sample t-test

```
data: .  
t = -0.56065, df = 24, p-value = 0.5802
```

```
alternative hypothesis: true mean is not equal to 37
95 percent confidence interval:
 36.80235 37.11321
sample estimates:
mean of x
 36.95778
```

Les résultats fournis ont une forme particulière qui est utilisée par de nombreuses fonctions de tests statistiques dans R. Ils méritent donc qu'on s'y attarde un peu.

Sur la première ligne, R nous confirme que nous avons bien réalisé un test de Student à un échantillon. La première ligne de résultats fournit la valeur du t calculé (ici, -0.56), le nombre de degrés de libertés (ici, $df = 24$), et la p -value (ici, 0.58 , soit une valeur supérieure à α). Cette première ligne contient donc tous les résultats du test qu'il conviendrait de rappeler dans un rapport. On devrait ainsi dire :

Au seuil α de 5%, le test de Student ne permet pas rejeter l'hypothèse nulle $\mu = 37$ ($t = -0.56$, $ddl = 24$, $p = 0.58$). Les données observées sont donc compatibles avec l'hypothèse selon laquelle la température corporelle moyenne des adultes en bonne santé vaut 37°C .

C'est de cette manière que vous devriez rapporter les résultats de ce test dans un compte-rendu ou un rapport à partir de maintenant.

Dans les résultats du test, la ligne suivante (**alternative hypothesis: ...**) **ne donne pas la conclusion du test**. Il s'agit simplement d'un rappel concernant l'hypothèse alternative qui a été utilisée pour réaliser le test. Ici, l'hypothèse alternative utilisée est une hypothèse bilatérale ($\mu \neq 37$). Nous verrons plus tard comment spécifier des hypothèses alternatives uni-latérales, même si la plupart du temps, mieux vaut s'abstenir de réaliser de tels tests (à moins bien sûr d'avoir une bonne raison de le faire).

Les résultats fournis ensuite concernent, non plus le test statistique à proprement parler, mais l'estimation. Ici, la moyenne de l'échantillon est fournie. Il s'agit de la meilleure estimation possible de la moyenne de la population : $\bar{x} = \hat{\mu} = 36.96$. Comme pour toutes les estimations,

cette valeur est entachée d'incertitude liée à la fluctuation d'échantillonnage. L'intervalle de confiance à 95% de cette estimation de moyenne est donc également fourni : [36.80;37.11]. Vous notez qu'il s'agit des mêmes valeurs que celles que nous avons calculées dans la Section ??.

Autrement dit, cet intervalle contient les valeurs les plus vraisemblables pour la véritable valeur de moyenne dans la population générale. Cela confirme bien que nous n'avons pas prouvé au sens strict que la moyenne de la population vaut 37°C. Nous avons en réalité montré que nous ne pouvions pas exclure que la moyenne de la population générale soit de 37°C. Puisque cette valeur est comprise dans l'intervalle de confiance, on ne peut donc pas l'exclure : nos données sont compatibles avec cette hypothèse. Mais beaucoup d'autres valeurs figurent aussi dans cet intervalle. Il est donc tout à fait possible que la moyenne soit en réalité différente de 37°C (par exemple, 36.9°C). Pour en être sûr, il faudrait probablement un échantillon de plus grande taille afin de limiter l'incertitude, d'augmenter la puissance statistique de notre test, et ainsi d'être en mesure de détecter des différences subtiles.

1.7 L'alternative non paramétrique

Si jamais les conditions d'application du test de Student à un échantillon n'étaient pas remplies, il faudrait alors réaliser son équivalent non paramétrique : le **test de Wilcoxon des rangs signés**. Ce test est moins puissant que son homologue paramétrique. On ne l'effectue donc que lorsque l'on n'a pas le choix :

```
wilcox.test(Temp_clean$temperature, mu = 37, conf.int = TRUE)
```

ou

```
wilcox.test(temperature ~ 1, mu = 37, conf.int = TRUE, data = Temp_clean)
```

ou encore

```
Temp_clean %>%  
  pull(temperature) %>%
```

```
wilcox.test(mu = 37, conf.int = TRUE)
```

Warning in wilcox.test.default(., mu = 37, conf.int = TRUE): impossible de calculer la p-value exacte avec des ex-aequos

Warning in wilcox.test.default(., mu = 37, conf.int = TRUE): impossible de calculer un intervalle de confiance exact avec des ex-aequos

Wilcoxon signed rank test with continuity correction

```
data: .  
V = 143, p-value = 0.6077  
alternative hypothesis: true location is not equal to 37  
95 percent confidence interval:  
 36.77780 37.11114  
sample estimates:  
(pseudo)median  
 36.94446
```

La syntaxe est identique à celle du test de Student à un échantillon à une exception près : l'ajout de l'argument `conf.int = TRUE` qui permet d'afficher la (pseudo)médiane de l'échantillon et son intervalle de confiance à 95%.

Les hypothèses nulles et alternatives de ce test sont les mêmes que celles du test de Student à un échantillon. En toute rigueur, on compare la médiane à une valeur théorique, et non la moyenne. Mais dans la pratique, la grande majorité des utilisateurs de ce test font l'amalgame entre moyenne et médiane. Ici, la conclusion correcte devrait donc être :

Au seuil α de 5%, on ne peut pas rejeter l'hypothèse nulle (test de Wilcoxon des rangs signés, $V = 143$, $\widehat{p} = 0.6077$). La médiane de la population ($\widehat{med} = 36.94$) n'est pas significativement différente de 37°C (IC 95% : [36.78; 37.11]).

Si les données ne suivent pas la loi Normale, la médiane est bien la métrique la plus intéressante puisque c'est elle qui nous renseigne sur la tendance centrale des données.

Enfin, les tests de Wilcoxon renvoient souvent des messages d'avertissement. Il ne s'agit que de ça : des avertissements. Tant que la p -value d'un test est éloignée de la valeur seuil α , cela n'a pas d'importance. Quand en revanche la p -value est très proche de α , les messages d'avertissement doivent vous alerter : il faut être très prudent face aux conclusions du test qui peuvent alors être assez “fragiles”.

1.8 Les notions d'erreur et de puissance statistique

Pour avoir le droit de réaliser un test paramétrique, il faut au préalable vérifier qu'un certain nombre de conditions sont vérifiées. Si ce n'est pas le cas, on réalise un équivalent non paramétrique. On peut alors se demander pourquoi ne pas se contenter de faire des tests non paramétrique systématiquement, sans s'embêter à faire des tests supplémentaires ou des tests paramétriques.

La raison est simple et elle est liée aux notions d'erreur et de puissance statistique.

Définitions

- **Erreur de type I** : notée α , c'est la probabilité de rejeter à tort l'hypothèse nulle. C'est donc la probabilité de rejeter H_0 alors qu'elle est vraie.
- **Erreur de type II** : notée β , c'est la probabilité d'accepter à tort l'hypothèse nulle. C'est donc la probabilité d'accepter H_0 alors qu'elle est fausse.
- **Puissance statistique** : notée $1 - \beta$, c'est la probabilité de rejeter l'hypothèse nulle à raison. C'est donc la probabilité de rejeter H_0 quand elle est réellement fausse.

À chaque fois que l'on réalise un test statistique, on commet nécessairement les 2 types d'erreurs α et β . On souhaite évidemment minimiser les erreurs, mais on ne peut malheureusement pas faire baisser les 2 en même temps. Faire baisser α (pour diminuer les faux positifs) conduit toujours à augmenter β (les faux négatifs). Faire baisser α revient en

effet à accepter plus souvent l'hypothèse nulle quand elle est vraie. Cela conduit inévitablement à accepter aussi plus souvent l'hypothèse nulle quand elle est fausse (et donc, à augmenter les faux négatifs).

Pour bien comprendre l'enjeu associé à ces erreurs, prenons l'exemple de notre système judiciaire. Lorsqu'un accusé est jugé, il est présumé innocent jusqu'à preuve du contraire. Le procès est l'équivalent d'un test statistique, avec :

- H_0 : l'accusé est innocent
- H_1 : l'accusé est coupable

Commettre une erreur de type I revient à condamner à tort l'accusé (on rejette à tort H_0), donc on condamne un innocent. À l'inverse, commettre une erreur de type II revient à libérer un coupable (accepter à tort H_0). Un système de justice plus strict condamnera un plus grand nombre d'accusés, qu'ils soient coupables ou non. Un système plus strict fera donc augmenter l'erreur de type I et baisser l'erreur de type II. À vous de voir ce que vous préférez : libérer plus de coupables, ou condamner plus d'innocents ?

En statistiques, la question est tranchée puisqu'on préfère maintenir l'erreur de type I à un niveau assez faible (à 5% ou moins), quitte à laisser augmenter l'erreur de type II (qui est considérée comme acceptable jusqu'à 20% environ). Toutefois, seule l'erreur de type I est sous notre contrôle. En effet, c'est nous qui la choisissons lorsque l'on fixe le seuil α de nos tests statistiques.

! À retenir

C'est vous qui fixez l'erreur de type I lorsque vous faites un test statistique. L'erreur de type I est le seuil α du test, que l'on fixe en général à 0,05 (soit 5%) dans le domaine des sciences du vivant.

Une fois que le seuil α est fixé, l'erreur β l'est aussi dans une certaine mesure. Mais on ne peut la connaître avec précision car elle dépend de beaucoup de choses, notamment la taille des échantillons dont on dispose, la variabilité des données, le type de test réalisé, etc. En général, **plus la taille de l'échantillon sera grande, plus l'erreur β sera faible, et donc plus la puissance sera élevée**. De même, par

rapport aux tests non paramétriques, les tests paramétriques permettent de minimiser l'erreur β et donc d'augmenter la puissance.

Puisque la puissance statistique vaut $1 - \beta$, cela revient à dire que les tests paramétriques sont plus puissants que les tests non paramétriques (parfois, beaucoup plus). Au contraire des erreurs de type I et II, la puissance est une grandeur que l'on souhaite maximiser. On aimerait en effet être capables de systématiquement rejeter H_0 quand elle est fausse. Nous avons vu plus haut que c'est hélas impossible. Mais choisir le bon test et la bonne procédure statistique permettent néanmoins d'augmenter la puissance, jusqu'à un certain point. C'est la raison pour laquelle on réalisera toujours un test paramétrique si les données dont on dispose le permettent (donc si les conditions d'application des tests paramétriques sont respectées). Et ce n'est qu'en dernier recours qu'on se tournera vers les tests non paramétriques, toujours moins puissants.

! Important

Un test paramétrique est toujours plus puissant que ses homologues non paramétriques. Avec un test paramétrique, il est donc plus probable de rejeter H_0 à raison qu'avec un test non paramétrique.

1.9 Bilan

Nous avons vu dans ce chapitre quelle est la procédure à suivre pour réaliser un test de comparaison de la moyenne d'une population à une valeur théorique :

1. examen préliminaire des données
2. calcul de statistiques descriptives
3. création de graphiques exploratoires
4. vérification des conditions d'application du test paramétrique
5. réalisation du test paramétrique ou non paramétrique selon l'issue de l'étape 4

Mais nous avons aussi abordé des notions statistiques essentielles pour la suite :

1. Les ingrédients indispensables pour réaliser un test statistique (les hypothèses nulle et alternative, la statistique du test et le seuil α).
2. La p -value et la décision du test.
3. Les erreurs de type I (α) et II (β).
4. La puissance statistique ($1 - \beta$) qui n'a rien à voir avec la notion de précision.
5. La notion de test paramétrique ou non paramétrique.

Assurez-vous d'avoir les idées claires sur toutes ces notions car elles sont absolument centrales pour ne pas faire/dire de bêtises lorsque l'on analyse des données.

1.10 Exercice d'application

Le fichier [Temperature2.csv](#) contient les données brutes d'une seconde étude similaire, réalisée à plus grande échelle. Importez ces données et analysez-les afin de vérifier si la température corporelle moyenne des adultes en bonne santé vaut bien 37°C . Comme toujours, avant de vous lancer dans la réalisation des tests statistiques, prenez le temps d'examiner vos données comme nous l'avons décrit dans la Section ?? et la Section ??, afin de savoir où vous allez, et de repérer les éventuelles données manquantes ou aberrantes. Enfin, interprétez les résultats à la lumière des notions que nous avons abordées ici (en particulier la notion de puissance statistique).

2 Comparaison de moyennes : deux échantillons appariés

2.1 Pré-requis

Pour ce 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 les mêmes packages que précédemment :

- 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 que vous pouvez dès maintenant télécharger dans votre répertoire de travail :

- [Autruches.csv](#)
- [Testosterone.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())
```

2.2 Contexte

On s'intéresse ici à la comparaison de 2 séries de données dont les observations sont liées 2 à 2. C'est par exemple le cas lorsque l'on fait subir un traitement à différents sujets et que l'on souhaite comparer les mesures obtenues avant et après le traitement.

Autrement dit, dans les plans d'expériences appariés, **les deux traitements ou modalités sont appliqués à chaque unité d'échantillonnage** : chaque sujet ou unité d'échantillonnage fournit plusieurs valeurs. Ça n'était pas le cas du chapitre précédent (@#sec-moy1) où chaque adulte n'avait fourni qu'une unique valeur de température.

Voici quelques exemples de situations qui devraient être traitées avec des tests sur données appariées :

- Comparaison de la masse de patients avant et après une hospitalisation.
- Comparaison de la diversité de peuplements de poissons dans des lacs avant et après contamination par des métaux lourds.
- Test des effets d'une crème solaire appliquée sur un bras de chaque volontaire alors que l'autre bras ne reçoit qu'un placebo.
- Test des effets du tabagisme dans un échantillon de fumeurs, dont chaque membre est comparé à un non fumeur choisi pour qu'il lui ressemble le plus possible en terme d'âge, de masse, d'origine ethnique et sociale, etc.