FLORENCE NICOLÈ • DAMIEN MARAGE
GUY LEMPÉRIÈRE • JORGE PAULO CANCELA DA FONSECA

# Guide pratique de biostatistique

## Choisir les bons outils pour analyser ses données

LICENCES, MASTERS,
CAPES DE SVT
AGRÉGATION SV-STU

- Cours complet
- Exemples et scripts R commentés
- Arbre décisionnel des tests

deboeck SUPÉRIEUR

# Guide pratique
## de biostatistique

## Dans la collection Licence Master Doctorat

***Biologie***

CORNEC J.-P., *La cellule eucaryote*

CORNEC J.-P., *Immunité des métazoaires. Aspects cellulaires, moléculaires, génétiques et évolutifs*

COUÉE Y., FONTAINE-POITOU L., GUILLAUME V., *Biologie et physiologie cellulaires et moléculaires*

GAUTHIER CLERC M., THOMAS F., *Écologie de la santé et biodiversité*

HERVÉ M., *Systématique animale. D'Aristote aux phylogénies moléculaires : histoire, concepts et méthodes de la classification*

GARNIER É., NAVAS M.-L., *Diversité fonctionnelle des plantes*

GUÉGAN J.-F., CHOISY M., *Introduction à l'épidémiologie intégrative des maladies infectieuses et parasitaires*

MAZLIAK P., *Le déterminisme de la floraison : contrôles génétiques et épigénétiques*

POINSOT D., HERVÉ M., LE GARFF B., CEILLIER M., *Diversité animale. Histoire, évolution et biologie des Métazoaires*

TANZARELLA S., *Perception et communication chez les animaux*

THOMAS F., GUÉGAN J.-F., RENAUD F., *Écologie et évolution des systèmes parasités*

## Chez le même éditeur

FORÊT R., *Dictionnaire des sciences de la vie*

MIQUEL P.-A, *Biologie du XXIe siècle : évolution des concepts fondateurs*

RAVEN P.H., JOHNSON G.B., LOSOS J.B., MASON K.A., SINGER S.R., *Biologie*. 5e éd.

RAVEN P.H., EVERT R.F., EICHHORN S.E., *Biologie végétale*. 3e éd.

SINGH CUNDY A., SHIN G., *Découvrir la biologie*. 2e éd.

THOMAS F., LEFEVRE T., RAYMOND M., *Biologie évolutive*. 2e éd.

ZIMMER C., *Introduction à l'évolution*

## Dans la collection Memento Sciences

AYMERIC J.-L., LEFRANC G., *Immunologie humaine*

GALAS S., DESCAMPS S., MARTINEZ A.-M., *Le cycle cellulaire*

GAUDRIAULT S., VINCENT R., *Génomique*

NGUYÊN V., FERRY N., *La reproduction des vertébrés*

VINCENT R., *Génétique moléculaire*

Florence **Nicolè**, Damien **Marage**, Guy **Lempérière**,
Jorge Paulo **Cancela Da Fonseca**

# Guide pratique de biostatistique

## Choisir les bons outils pour analyser ses données

biologie

LICENCES, MASTERS,
CAPES DE SVT
AGRÉGATION SV-STU

deboeck SUPÉRIEUR

Pour toute information sur notre fonds et les nouveautés dans votre domaine de spécialisation, consultez notre site web: **www.deboecksuperieur.com**


Rue du Bosquet, 7, 1348 Louvain-la-Neuve



Dépôt légal :
Bibliothèque nationale, Paris : septembre 2022
Bibliothèque royale de Belgique, Bruxelles : 2022/13647/138 ISBN 978-2-8073-3167-9

# Sommaire

# Remerciements

Le manuscrit de cet ouvrage a été très significativement amélioré grâce à la relecture approfondie et toujours perspicace de Frédéric Gosselin, ingénieur général des ponts, des eaux et des forêts, chercheur à l'INRAe en écologie forestière et biométrie. Les auteurs lui expriment leur plus profonde gratitude.

Nous témoignons également notre plus profonde reconnaissance à Aurélien Besnard, directeur d'études à l'École Pratique des Hautes Études, pour la préface de cet ouvrage.

Nous tenons enfin à remercier tout particulièrement Fabrice Chrétien des éditions De Boeck Supérieur de nous avoir permis de mener jusqu'à son terme ce guide. Le climat de confiance et amical qu'il a su instaurer, y est pour beaucoup.

# Préface

L'expérience naturaliste et l'observation de la nature sont à la base de l'évolution des disciplines scientifiques que sont la biologie et l'écologie. Longtemps cantonnées à des démarches observationnelles et descriptives, ces disciplines se sont imposées, sur les dernières décennies, comme de véritables sciences quantitatives, fondées sur des théories, des concepts et des démarches hypothético-déductives. Une telle évolution, qui leur permet aujourd'hui de faire autorité et par exemple de répondre à des questions de société, d'alimenter les réflexions sur les politiques publiques à mettre en œuvre, s'est cependant accompagnée d'une complexité croissante dans les outils analytiques mobilisés. En particulier, la Statistique est désormais au cœur de ces disciplines car elle permet une quantification des incertitudes, mais aussi de tester formellement les hypothèses de travail. Ce passage nécessaire de l'observation et de la description à la formulation mathématique/statistique n'est cependant pas une évidence pour nombre d'étudiants et professionnels en biologie et écologie. La statistique renvoie à la notion de modèle, et donc de caricature, s'éloignant de l'approche sensible qui est souvent à la base de la passion pour la biologie et l'écologie. Pourtant, « faire parler des données », souvent fortement bruitées, issues de l'observation est aussi une grande source de satisfaction. Formuler des hypothèses issues de sa propre expérience naturaliste, de ses constats de terrain, les formaliser et surtout les valider est, quoi qu'on en dise, une forme de consécration de son expérience de terrain. Alors pourquoi la Statistique est-elle la bête noire de beaucoup d'étudiants et de professionnels en biologie/écologie ? Cela tient souvent du mythe de la déconnexion entre statistiques et expérience de terrain, mythe largement entretenu par un enseignement universitaire longtemps très théorique de cette discipline pour les biologistes/écologues. De grands progrès ont cependant été faits ces dernières années dans l'enseignement de la Statistique pour les biologistes/écologues qui ont partiellement fait tomber ce mythe. Il reste malgré tout que la Statistique est un domaine très vaste, de nombreux chercheurs statisticiens continuent de développer de nouvelles méthodes pour gérer des problèmes de plus en plus pointus. Les étudiants peuvent de fait se retrouver rapidement démunis face à l'apparente immensité de la tâche de se former à la Statistique appliquée.

Florence Nicolè, Damien Marage, Guy Lempérière et Jorge Cancela Da Fonseca (1926-2011) signent ici un ouvrage didactique, court, accessible au plus grand nombre que cela soit des étudiants en biologie ou écologie mais aussi des profes-

sionnels qui chercheraient à se former aux statistiques. Les auteurs ont fait le choix judicieux de se concentrer sur les méthodes les plus couramment utilisées, des méthodes qui suffiront pour traiter une vaste majorité de situations rencontrées en biologie/écologie. Ils évitent ainsi le piège de la volonté d'exhaustivité (jamais atteinte) laissant souvent le sentiment d'inaccessibilité aux lecteurs naïfs. La volonté des auteurs de produire un ouvrage court ne tombe pas, par ailleurs, dans le travers du simple guide de programmation des tests. Cet ouvrage balaie en effet tous les points importants d'une démarche rigoureuse : formuler des hypothèses de travail, définir des plans d'échantillonnage, explorer ses données, choisir et conduire des tests adaptés, etc. Il fait ainsi la part belle à la science, rappelant que si la Statistique est une discipline de recherche, il est crucial de ne pas oublier qu'elle doit être au service de la question biologique/écologique dans nos domaines. Par ce focus sur la démarche générale à laquelle s'ajoute des parties techniques pour mettre en œuvre concrètement les tests sous R et interpréter leurs sorties, cet ouvrage vient combler un gap qu'il y a entre d'un côté des ouvrages très théoriques faisant l'impasse sur la mise en œuvre concrète des analyses et d'un autre côté des guides très pratico-pratiques de mise en œuvre, sous R par exemple, faisant l'impasse sur les concepts et la démarche. En ce sens le titre de l'ouvrage est révélateur de cet objectif : véritable « guide pratique », il fait la part belle à la compréhension (« comprendre pour faire les bons choix »).

Cet ouvrage est donc à mettre entre toutes les mains de personnes souhaitant s'initier aux statistiques en relative autonomie.

Aurélien Besnard

Directeur d'Étude de l'École Pratique
des Hautes Études

UMR5175 Centre d'Écologie Fonctionnelle
et Évolutive, Montpellier

# Avant-propos

« À Jorge Paulo Cancela da Fonseca (1926-2011) et Simon Leather (1955-2021) *in memoriam* ».

Ce guide n'est pas et ne se veut pas un énième manuel de statistiques à l'usage des biologistes. Il est conçu pour être à destination des étudiants ou praticiens qui utilisent les statistiques dans le cadre de leurs projets ou de leurs travaux personnels de biologie et d'écologie. Il est le résultat de la synthèse de cours, de travaux de thèses, de publications des différents auteurs et essaye, modestement, de guider les utilisateurs vers les meilleurs choix possibles de traitements de leurs données de terrain ou de laboratoire.

Il s'inspire largement des travaux pionniers de J.-P. Cancela da Fonseca (CNRS) qui a, dans les années 1980, développé l'outil statistique en biologie du sol, des travaux de D. Marage (Université de Besançon), K. Day (Université d'Ulster), S. Leather (Université Harper Adams) et G. Lempérière (LECA Université de Grenoble) en écologie forestière, de F. Nicolè (Université de Saint Étienne) en écologie végétale, de Jinliang Liu (Université de Wenzhou, Chine) en sciences de l'environnement et du remarquable guide de C. Dytham, *Choosing and Using Statistics, A Biologist's Guide*. Il est structuré autour de 10 chapitres, d'un arbre de décision de tests et de liens et de scripts permettant de les réaliser avec le logiciel R.

www.lienmini.fr/33167-scriptR

Nous nous sommes aussi efforcés d'intégrer le minimum d'équations dans le texte afin d'en faciliter la lecture et avons essayé de limiter au maximum l'utilisation du jargon statistique. Pour avoir trop souvent vu des étudiants hagards et désemparés après avoir subi des cours de statistiques incertains, voire chancelants, nous avons délibérément pris le parti de mener les utilisateurs vers la méthode la plus appropriée pour comprendre au mieux les tests et interpréter les résultats le plus correctement possible. Si un message devait ressortir de ce guide, il serait de réfléchir, d'anticiper, dès le début du travail de recherche, sur les outils statistiques qui pourraient être utilisés avant de collecter des données. Cela éviterait un gaspillage de temps, permettrait d'obtenir des résultats exacts et précis et éviterait de tomber

dans le piège voulant que l'on ne pense aux statistiques qu'en dernier lieu pour traiter les données.

Enfin, nous sommes loin de vouloir prétendre mettre le monde en équations, et nous agrémenterons notre propos en laissant les lecteurs et utilisateurs méditer sur ces quelques citations concernant les statistiques, qui, si elles nous procurent souvent des sueurs froides, ne manquent parfois pas de poésie comme nous le rappelle Eugène Labiche : « la statistique, madame, est une science moderne et positive. Elle met en lumière les faits les plus obscurs. Ainsi, dernièrement, grâce à des recherches laborieuses, nous sommes arrivés à connaître le nombre exact de veuves qui sont passées sur le Pont Neuf pendant le cours de l'année 1860 » (E. Labiche, *Les Vivacités du capitaine Tic* in S. Tesson, *La Panthère des neiges*). Autre exemple, Alphonse Allais, en 1890, ajoutera que « la statistique a démontré que la mortalité dans l'armée augmente sensiblement en temps de guerre », on peut ajouter celle des grands ongulés et des randonneurs en temps de chasse. Les statistiques peuvent même parfois être teintées d'un brin d'humour : « Voici les chiffres communiqués par les services de la statistique et intéressant la période comprise entre le 2 juillet et le 4 septembre : 543285 ; 6282826 ; 1285938743 ; 601 ; 602 ; 603 ; 604 ; 605 ; 106 ; 206 ; 206 ; 406 ; 506 ; 983 ; 882 ; 780 ; 680 ; 579. Nous ne savons pas du tout à quoi se rapportent ces chiffres, mais nous sommes heureux de les communiquer à nos lecteurs qui auront ainsi toute latitude de les adapter suivant leur goût ou leur appréciation… » (Pierre Dac, *L'Os à Moelle*, Juillard, Paris, 1963). Quant aux chiffres, selon plusieurs sources bien informées, ils sont, parait-il, « aux analystes ce que les lampadaires sont aux ivrognes : ils fournissent bien plus un appui qu'un éclairage. Ils sont comme les gens. Si on les torture assez, on peut leur faire dire n'importe quoi… étonnant non… ? »

Plus sérieusement, et contrairement à l'affirmation des frères Goncourt qui faisaient des statistiques la première des sciences inexactes, les statistiques et les probabilités constituent des outils puissants et indispensables aux sciences de la vie lorsqu'elles sont correctement utilisées et accompagnées d'analyses pertinentes. Nous espérons que ce guide contribuera à une meilleure compréhension et à une utilisation adaptée de ce qui est et restera une science exacte.

Bon courage et bonne lecture à toutes et tous,

Les auteurs

Chapitre 1

# Les dix étapes pour une analyse de données réussie et valider ces compétences

Ce premier chapitre est comme la colonne vertébrale de cet ouvrage. Il structure et articule les chapitres, et constitue les bases d'une analyse de données réussie. La séquence très simple de dix étapes que nous vous proposons ici devrait toujours guider vos analyses de données (figure 1). Lorsqu'elle est correctement suivie, elle permet d'intégrer les statistiques dans votre démarche de recherche et d'y apporter robustesse et fiabilité. Les outils statistiques que vous utiliserez devront ainsi être clairement explicités et identifiés très tôt dans votre démarche et non à la fin du processus comme c'est trop souvent le cas.

## 1. Réussir ses analyses de données en dix étapes

### Étape 1 :

Définissez une problématique que vous souhaitez traiter. Les notions fondamentales en statistiques vous seront nécessaires (*cf.* chapitre 2). Vous pouvez choisir de répliquer une étude déjà menée (reproductibilité et incertitudes, *cf.* chapitre 3) ou de tester une nouvelle hypothèse. Cela implique d'avoir mené sérieusement un état de l'art des connaissances qui existe sur le sujet.

### Étape 2 :

Déterminez et formulez une ou plusieurs hypothèses de travail (*cf.* chapitres 2 et 3 pour vous guider).

### Étape 3 :

Choisissez les dispositifs de mesures, le ou les plans d'échantillonnage, le ou les plans d'expérimentation qui vous permettront de tester vos hypothèses (*cf.* chapitre 3). Identifiez clairement la nature de vos variables (quantitative, qualitative, ordinale…), lesquelles seront explicatives ou à expliquer. Identifiez le format de tableau de données adapté à chaque test.

### Étape 4 :

Collectez un petit jeu de données « test » ou jeu de données d'apprentissage. Vous pourrez ainsi explorer graphiquement celui-ci (chapitre 5) et également déceler sa loi de distribution (chapitre 4). Cela peut paraître étrange mais cette approche permettra de rendre concret le protocole expérimental ou le plan d'échantillonnage proposé. Ce processus essai-erreur peut révéler des défauts ou des faiblesses dans le protocole et permet de les corriger et d'économiser énormément de temps et d'efforts (*cf.* chapitre 3). S'il est impossible de collecter des données avant de mener l'expérimentation complète, vous pouvez créer un jeu de données fictif basé sur des valeurs approximatives que vous vous attendez à obtenir. Vous pouvez par exemple tester votre matériel sur des échantillons factices.

### Étape 5 :

Utilisez l'arbre de décision présenté au chapitre 6 pour vous guider dans la recherche du ou des tests les plus appropriés pour répondre à votre objectif et statuer sur vos hypothèses.

### Étape 6 :

Effectuez le ou les tests sur votre jeu de données fictif ou d'apprentissage. Vous trouverez dans les chapitres 7 à 10 l'ensemble des tests et analyses réalisés sous R avec leurs scripts et l'interprétation des résultats. Vérifiez que ces analyses vous permettent d'apporter une réponse à votre objectif.

## Étape 7 :

Si l'étape 6 ne vous permet pas de répondre à votre objectif et de statuer sur vos hypothèses, vous devez revenir à l'étape 3 (ou 2) et réfléchir de nouveau à votre protocole. Si vous rencontrez des problèmes, revenez aux étapes 3 (ou 2). Il est possible de mener des tests de puissance sur le jeu de données d'apprentissage pour affiner la taille d'échantillonnage à prévoir. Quand vous aurez levé les problèmes méthodologiques et pratiques, que votre protocole parait fonctionnel, vous pourrez procéder à la collecte des données réelles.

## Étape 8 :

Stockez et organisez les données pour les retrouver, les conserver et en faciliter l'accès et la gestion (avec un gestionnaire de fichiers, un espace de stockage en ligne, des classeurs, des bases de données, un système d'information…). Nettoyez les données en gardant la traçabilité des modifications effectuées.

## Étape 9 :

Visualisez les données graphiquement (*cf.* chapitre 5) et revenez au nettoyage des données (étape 8) si nécessaire. Au besoin, envisager des transformations de variables pour valider les prérequis de certains tests. Réalisez-les sur les données récoltées.

## Étape 10 :

Interprétez les résultats, statuez sur les hypothèses de travail de l'étape 2 et répondez à l'objectif initial. Discutez les résultats en fonction de l'état de l'art de l'étape 1. Créez des figures pour illustrer les résultats et savoir adapter le rendu en fonction de l'usage envisagé.

Nous ne pouvons que vous conseiller vivement de suivre scrupuleusement cette séquence de dix étapes fondamentales pour le traitement de vos données (figure 1). Elle vous aidera à récolter, classer, traiter et interpréter vos données de manière pertinente.

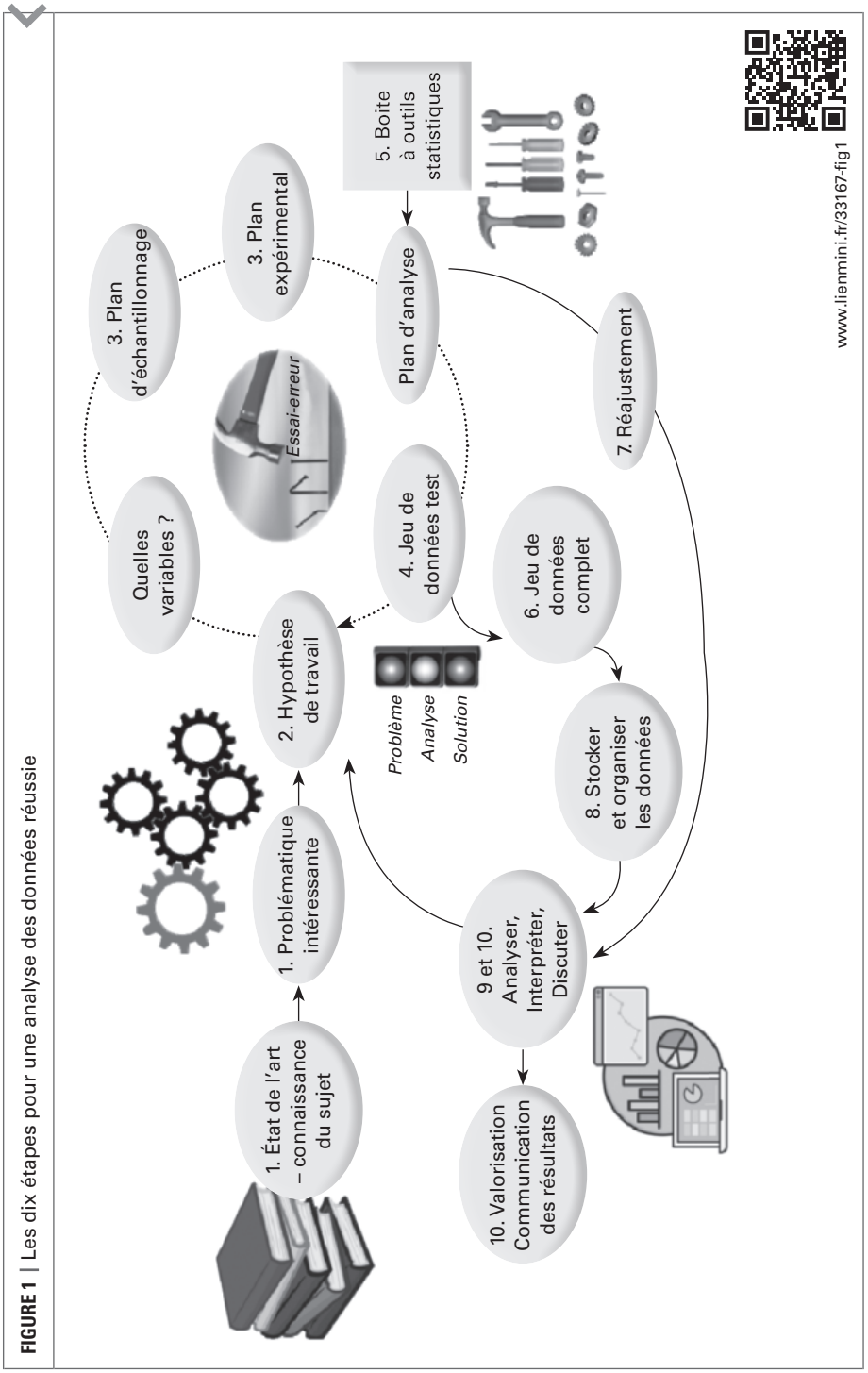



**FIGURE 1** | Les dix étapes pour une analyse des données réussie

## 2. Guide pratique pour l'analyse des données et certification aux outils numériques Pix : une approche par compétences

La certification Pix, qui a remplacé la certification C2i, est désormais la référence européenne en matière de compétences numériques pour l'enseignement supérieur et dans tous les milieux socio-professionnels.

Pix est une plateforme du service public en ligne (https://pix.fr), pour évaluer, développer et certifier ses compétences numériques tout au long de la vie. Chaque personne peut évaluer son niveau allant du niveau débutant (1) au niveau expert (8), dans 16 compétences numériques réparties en 5 grands domaines.

Comme vous le comprendrez à la lecture de ce guide pratique, l'analyse des données ne se limite pas à la connaissance des tests statistiques. Elle englobe des savoirs, des savoir-faire techniques et méthodologiques. Citons par exemple : savoir faire un état de l'art, savoir établir un protocole expérimental robuste pour répondre à une question bien identifiée, disposer d'un bon sens critique pour savoir discuter les résultats et le protocole, savoir présenter ses résultats en adaptant son discours à son public (tableau 1).

L'approche par compétences remplace progressivement l'approche par la connaissance dans l'enseignement supérieur. Dans les 5 blocs transversaux du Répertoire national des certifications professionnelles (RNCP) des compétences ci-dessous, les compétences abordées dans ce guide pratique d'analyse des données apparaissent en grisé.

**TABLEAU 1** | Les cinq blocs transversaux du Répertoire national des certifications professionnelles (RNCP) appliqués à notre guide

| | Intitulé du bloc de compétences | Compétences |
|---|---|---|
| **Bloc 1** | Usages digitaux et numériques | Utiliser les outils numériques de référence et les règles de sécurité informatique pour acquérir, traiter, produire et diffuser de l'information ainsi que pour collaborer en interne et en externe |
| **Bloc 2** | Exploitation de données à des fins d'analyse | Identifier, sélectionner et analyser avec esprit critique diverses ressources dans son domaine de spécialité pour documenter un sujet et synthétiser ces données en vue de leur exploitation |
| | | Analyser et synthétiser des données en vue de leur exploitation |
| | | Développer une argumentation avec esprit critique |

→

| | Intitulé du bloc de compétences | Compétences |
|---|---|---|
| **Bloc 3** | Expressions et communications écrites et orales. | Se servir aisément des différents registres d'expression écrite et orale de la langue française |
| | | Communiquer par oral et par écrit, de façon claire et non ambiguë, dans au moins une langue étrangère |
| **Bloc 4** | Positionnement vis-à-vis d'un champ professionnel | Identifier et situer les champs professionnels potentiellement en relation avec les acquis de la mention ainsi que les parcours possibles pour y accéder |
| | | Caractériser et valoriser son identité, ses compétences et son projet professionnel en fonction d'un contexte |
| | | Identifier le processus de production, de diffusion et de valorisation des savoirs |
| **Bloc 5** | Action en responsabilité au sein d'une organisation professionnelle | Situer son rôle et sa mission au sein d'une organisation pour s'adapter et prendre des initiatives |
| | | Respecter les principes d'éthique, de déontologie et de responsabilité environnementale |
| | | Travailler en équipe et en réseau ainsi qu'en autonomie et responsabilité au service d'un projet |
| | | Analyser ses actions en situation professionnelle, s'autoévaluer pour améliorer sa pratique |

En vous formant à l'analyse des données à travers la lecture de cet ouvrage, vous pourrez renforcer ou acquérir des compétences de la certification Pix. Nous listons ci-dessous l'ensemble des compétences Pix et nous avons matérialisé en *italique* les compétences développées dans le cadre de l'analyse de vos données.

## 2.1. Informations et données

- *Mener une recherche et une veille d'information pour répondre à un besoin d'information et se tenir au courant de l'actualité d'un sujet ;*
- *Stocker et organiser des données pour les retrouver, les conserver et en faciliter l'accès et la gestion ;*
- *Appliquer des traitements à des données pour les analyser et les interpréter.*

## 2.2. Communication et collaboration

- Interagir avec des individus et de petits groupes pour échanger dans divers contextes liés à la vie privée ou à une activité professionnelle, de façon ponctuelle et récurrente ;

- Partager et publier des informations et des contenus pour communiquer ses propres productions ou opinions, relayer celles des autres en contexte de communication publique ;
- *Collaborer dans un groupe pour réaliser un projet, co-produire des ressources, des connaissances, des données, et pour apprendre ;*
- S'insérer dans le monde numérique.

## 2.3. Création de contenu

- *Produire des documents à contenu majoritairement textuel pour communiquer des idées, rendre compte et valoriser ses travaux ;*
- *Développer des documents à contenu multimédia pour créer ses propres productions multimédias, enrichir ses créations majoritairement textuelles ou créer une œuvre transformative ;*
- *Adapter des documents de tous types en fonction de l'usage envisagé et maîtriser l'usage des licences pour permettre, faciliter et encadrer l'utilisation dans divers contextes ;*
- Écrire des programmes et des algorithmes pour répondre à un besoin et pour développer un contenu riche.

## 2.4. Protection des données et sécurité

- *Sécuriser les équipements, les communications et les données pour se prémunir contre les attaques, pièges, désagréments et incidents susceptibles de nuire au bon fonctionnement des matériels, logiciels, sites internet, et de compromettre les transactions et les données ;*
- Maîtriser ses traces et gérer les données personnelles pour protéger sa vie privée et celle des autres, et adopter une pratique éclairée ;
- Prévenir et limiter les risques générés par le numérique sur la santé, le bien-être et l'environnement mais aussi tirer parti de ses potentialités pour favoriser le développement personnel, le soin, l'inclusion dans la société et la qualité des conditions de vie, pour soi et pour les autres.

## 2.5. Environnement numérique

*Installer, configurer et enrichir un environnement numérique (matériels, outils, services) pour disposer d'un cadre adapté aux activités menées, à leur contexte d'exercice ou à des valeurs.*

Chapitre 2

# Les fondamentaux

L'objectif principal de ce chapitre consiste à introduire en termes assez généraux quelques-unes des notions de base sur la collecte des données et leur analyse. Tous les concepts abordés dans ce chapitre feront l'objet de développements plus détaillés dans les chapitres suivants et vous les trouverez également dans la plupart des ouvrages de statistiques qui abordent les plans d'expérimentation et d'échantillonnage.

En biologie et en écologie, nous travaillons souvent sur des individus ou des groupes d'individus, depuis des groupes de cellules par exemple, jusqu'à des populations d'insectes ou de plantes. Dans la plupart des cas, si l'on ne peut pas mesurer chaque individu, on est amené à utiliser un jeu d'individus du groupe ou de la population que l'on désigne par l'**échantillon**. On peut ainsi analyser des questions qui se posent à propos de nos groupes en formulant des hypothèses. Une question toute simple pourrait être : l'espèce A est-elle plus grosse que l'espèce B ? On pourrait facilement y répondre si l'on disposait des données pour tous les individus du groupe. Si nous n'avons qu'un échantillon du groupe, nous sommes amenés à extrapoler à l'ensemble du groupe. C'est le but de la statistique.

Les notions abordées dans ce chapitre constituent un prérequis pour bien utiliser l'arbre de décision qui est donné au chapitre 6.

## 1. Variables, individus et données

**Les variables constituent le matériel de base des statistiques.** On parle aussi de caractères ou d'observations statistiques. Ce sont tous les éléments qui peuvent être relevés lors d'une expérimentation et dont on pense qu'ils peuvent aider à répondre

à la question que l'on se pose. Cela peut aller d'une mesure chiffrée précise, qui évalue directement le processus qui nous intéresse et s'élargit à tous les éléments internes ou externes qui peuvent influencer cette mesure.

On souhaite par exemple analyser les différences d'abondances des communautés d'acariens sur 2 champs différents : une rizière et un champ abandonné (jeu de données exemple 1). On pense que la période à laquelle on va effectuer la mesure a une influence importante car certaines espèces ne survivent pas au froid ou au chaud. On décide donc de comparer la composition en espèces des communautés pour chaque mois de l'année, et sur chacun des deux sites indépendamment. On établit un protocole standardisé qui permet à chaque acarien présent d'avoir la même probabilité d'être dénombré, quel que soit le mois ou le site. Pour chaque acarien prélevé, on va déterminer son espèce. D'autres éléments internes peuvent être associés tels que le stade (juvénile ou adulte) ou le sexe, s'il est possible de le déterminer. Des éléments externes pouvant influencer les résultats peuvent être notés tels que l'observateur ou l'échantillonneur, la météorologie du jour de prélèvement ou des caractéristiques du sol. Toutes ces informations ne seront pas forcément utiles par la suite mais elles peuvent permettre de comprendre un résultat inattendu ou de pointer un facteur que l'on ne pensait pas important au premier abord. Quelle que soit l'investigation, nous conseillons vivement d'avoir toujours un support pour noter un maximum d'éléments sur le déroulement de l'expérimentation (cahier de terrain ou de manipulation, tablette numérique). Ces éléments seront conservés précieusement jusqu'à la valorisation des résultats.

**Un individu** ou unité statistique est un élément de base constitutif de la population statistique étudiée. C'est l'objet sur lequel on effectue les observations ou les mesures. Le plus souvent l'individu est repéré dans l'espace (placettes, commune…) et dans le temps. Il peut être un animal, un végétal, un humain, un objet, etc.

Exemple : des plantes comparées pour leur production d'huile essentielle, des acariens, des arbres.

**Une population statistique** est un ensemble d'individus sur lequel porte l'étude. C'est l'ensemble des objets ou des êtres vivants qui sont étudiés. Nos hypothèses de travail seront testées sur cette population et il est possible que le résultat ne soit pas le même sur une autre population.

Exemple : un champ expérimental de plantes, des sites pour lesquels on étudie la composition en acariens ou la banque de graines du sol.

**Une donnée** est la réalisation d'une variable mesurée sur un individu statistique. On suppose la mesure correcte et le choix de l'individu aléatoire car nous souhaitons tester des hypothèses dessus (on parle d'inférence statistique). L'ensemble des

données constitue le jeu de données. Tout élément qui est observé ou mesuré sur les individus d'une population statistique constitue une variable statistique. Clarifions encore quelques éléments de vocabulaire :

On divise généralement les variables en quatre grands types :

1. Les variables se rapportant à des catégories sans véritable échelle et pour lesquelles il est impossible d'établir un ordre. On parle de **variables qualitatives nominales**. *Chaque espèce d'acariens constitue une catégorie et il est impossible de dire qui est supérieur à qui.*
2. Les variables non mesurables mais pour lesquelles il est possible d'établir un ordre. On parle de **variables qualitatives ordinales**. *Par exemple, les mois sont des données qualitatives qui peuvent être ordonnées chronologiquement.*
3. Les variables peuvent être **quantitatives discrètes** lorsque les valeurs sont des nombres entiers issus d'un dénombrement. *Par exemple, le nombre d'acariens d'une espèce donnée est forcément un nombre entier.* Les variables discrètes ou discontinues ont un nombre limité de valeurs entre deux bornes.
4. Enfin, les variables peuvent être **quantitatives continues** lorsque les valeurs sont des nombres décimaux ; toutes les valeurs sont théoriquement possibles et uniquement limitées par les dispositifs de mesures (concentrations, longueurs…). *Par exemple, le pH du sol où ont été effectués les prélèvements d'acariens sera une valeur quantitative continue s'il est mesuré avec un pH-mètre.* Ce type de variable a théoriquement un nombre infini de valeurs entre deux bornes. En pratique, l'**exactitude** des mesures n'est pas parfaite car elle est limitée par l'observateur et l'équipement utilisé. Il y aura alors un nombre limité de valeurs possibles entre les deux bornes. Les longueurs, les surfaces et le poids sont des variables continues. On peut effectuer un découpage en classes pour traiter ces données.

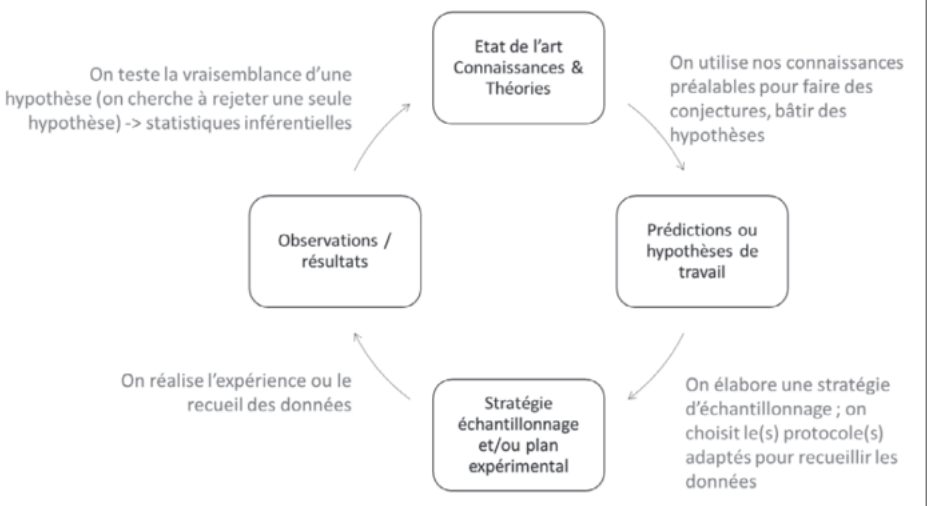
Il est très important de clarifier la nature des variables que vous voulez utiliser puis celles que vous avez réellement mesurées. *Par exemple, le pH-mètre de terrain est tombé en panne et nous n'avons pu utiliser que la réaction colorimétrique d'un papier pH. Nous avons alors comme information, non plus une valeur quantitative continue comme nous l'attendions, mais une observation catégorielle (acide, neutre ou basique).*

Le type de variables recueillies conditionne le type de tests statistiques à utiliser. Il est donc primordial de réfléchir en amont de la collecte des données sur les tests qui pourront être menés avec ces variables. Ne pas le faire, c'est prendre le risque de se retrouver face à un jeu de données qui ne peut être traité statistiquement pour répondre à l'objectif.

# 2. Méthode scientifique hypothético-déductive et tests d'hypothèses

La **méthode scientifique** désigne le processus par lequel des connaissances scientifiques sont produites. Il existe de nombreuses méthodes scientifiques mais l'une des plus utilisées est la méthode hypothético-déductive. Elle consiste à formuler des hypothèses de travail sur la base de nos connaissances. Ce processus se veut objectif mais en réalité, nos intuitions et celles de ceux qui ont travaillé sur le sujet auparavant influencent nos prédictions et nos hypothèses de travail. Pour tester ces prédictions, une stratégie d'échantillonnage ou un plan expérimental doivent être établis pour ainsi recueillir les données. Ces observations vont permettre d'appuyer ou de réfuter les hypothèses (figure 1).

**FIGURE 1** | Une démarche hypothético-déductive en analyse des données




Une **hypothèse de travail** consiste à annoncer le résultat d'une expérimentation avant d'avoir effectué les mesures et mené les observations sur la base de connaissances. On cherche donc à prédire ce que l'on attend quand on réalisera l'expérimentation.

Une hypothèse de travail peut être une déclaration du type « Les populations d'acariens d'un champ de riz sont plus abondantes que celles d'un champ abandonné ».

# Guide pratique de biostatistique

## Choisir les bons outils pour analyser ses données

**Ce guide pratique est destiné aux étudiants ou praticiens qui utilisent les statistiques dans le cadre de leurs projets ou de leurs travaux personnels de biologie et d'écologie. La statistique est en effet au cœur de ces disciplines, car elle permet de tester formellement des hypothèses de travail et constitue un passage obligé de l'observation et de la description à la formulation mathématique et au traitement des données.**

En dix chapitres, il couvre un ensemble de méthodes et de tests couramment utilisés et insiste sur les points cruciaux d'une démarche scientifique au service des questions de biologie et d'écologie. Un arbre de décision est proposé pour effectuer les bons choix de tests. Les auteurs proposent également de nombreux exemples tirés de la littérature scientifique, de leurs travaux en écologie végétale, écologie du sol et écologie forestière. Des scripts permettant de les réaliser sous le logiciel R sont également insérés tout au long du guide.


### LES AUTEURS

**Florence Nicolè** est Maître de Conférences à l'Université Jean Monnet de Saint-Étienne. Elle est chercheuse au laboratoire de biotechnologies végétales appliquées aux plantes aromatiques et médicinales (LBVPAM UMR CNRS 5079). Elle est Présidente du Conservatoire National des Plantes à Parfum, Médicinales, Aromatiques et Industrielles (CNPMAI).

**Damien Marage** est Professeur de géographie à l'Université de Franche Comté. Il conduit ses recherches sur la prise en charge du vivant dans les territoires au sein du laboratoire ThéMA UMR 6049 CNRS à Besançon.

**Guy Lempérière** est Docteur ès Sciences. Il a été Professeur Associé à l'Université de Grenoble-Alpes et Chercheur au LECA (Laboratoire d'Écologie Alpine, UMR CNRS 5553) puis Directeur de Recherche à l'IRD.

**Jorge Paulo Cancela da Fonseca** (1926-2011) Docteur ès Sciences a été Directeur de Recherche au CNRS. Ses travaux en écologie du sol au sein du Laboratoire d'Écologie Forestière de Fontainebleau (Université Paris 7) lui ont permis de développer des outils statistiques puissants et performants qui lui ont valu la médaille de bronze du CNRS.


### LES PLUS

- **Des jeux de données, les scripts R et également des liens (avec QR codes) pour accéder aux figures en couleur**
- **Nombreux exemples d'application sous R**
- **Arbre décisionnel**


Conception graphique : Primo&Primo®

ISBN : 9782-8073-3167-9

9 782807 331679

deboeck SUPÉRIEUR

www.deboecksuperieur.com