Statistique descriptive 1
RES 1-05

Maeva Paradis et Paul-Marie Grollemund

Table des matières

CHAPITRE 1

Introduction

La lecture de ce chapitre a pour objectif d'introduire la ressource pédagogique "Statistique descriptive 1". Pour y aller progressivement, ce chapitre commence en précisant la signification des termes importants de cette ressource, ce qui permet de percevoir ce qu'il y a à avoir en tête concernant le contexte de ce cours, mais aussi qui permet de comprendre dans quel cadre s'insèrent les notions du cours.

Dans ce chapitre, il est tout d'abord donné en section 1.1 une explication de ce qu'est la statistique dans les grandes lignes. De plus, la section 1.2 correspond à un énoncé complet de ce qu'est la démarche scientifique à suivre dans le cadre d'une analyse statistique. Cela permet en particulier de comprendre à quoi correspond la statistique descriptive, et comment elle s'intègre en tant qu'étape dans un processus global, qui est l'analyse statistique. Ensuite, afin de se fixer les idées, la section 1.3 décrit un parallèle entre l'analyse statistique et un exemple parlant pour tous. La section 1.4 a pour objectif d'introduire le sous-contexte particulier de la statistique descriptive, ainsi que ces enjeux. Pour spécifier le propos de cette ressource pédagogique, la section 1.5 précise les objectifs de cette ressource en termes de notions à comprendre et de compétences à maîtriser. Pour finir ce chapitre, une mise en pratique d'une étude incluant des éléments de statistique descriptive est présentée en section 1.6. Il s'agit d'une application concernant l'impact du Covid19 sur le déroulement d'un championnat de football, de sorte à percevoir ce à quoi ressemble l'approche statistique en pratique, jusqu'à la phase d'une analyse descriptive.

Dans ce document, à la suite de ce chapitre, le chapitre 2 correspond à une première moitié de cette ressource et est dédiée aux outils de statistique descriptive univariée (pour traiter isolément une variable) et le chapitre 3 est une seconde moitié, qui concerne les outils de statistique descriptive bivariée (pour traiter des paires de variables).

Table des matières de ce chapitre

1.1 Une vision globale et rapide de la statistique

Une première question qu'un étudiant peut se poser en commençant cette formation est de savoir : "*Mais les stats c'est quoi en fait ?*" La question est légitime puisque cette discipline est assez légèrement traitée au lycée. Bien qu'elle soit au programme, elle est incluse dans le cadre d'un cours de mathématiques, qui contient un grand nombre de sous-disciplines, et qui a tendance à faire la part belle à d'autres contenus, parmi lesquelles l'analyse, la géométrie, l'algèbre, les probabilités ou encore l'algorithmique. En dehors du contexte scolaire, la statistique peut aussi être présente dans le vocabulaire de la vie de tout les jours, lorsqu'on parle des statistiques du match de la veille, des statistiques démographiques concernant la population française, des traitements statistiques en

science ou encore avec les médias d'information pour qui "la stat fait parler les chiffres". Autant de visions dont un étudiant a pu faire l'expérience concernant ce que serait la statistique, mais qui peuvent laisser dans un flou concernant ce qu'est réellement la statistique.

Pour éclaircir ce qu'il y a à savoir sur la statistique, voici ce qu'il faut avoir en tête. La statistique est la science de l'analyse des données sous un angle mathématique. Les données, qu'on dénomme aussi observations ou mesures, correspondent à une perception d'une réalité qu'on souhaite étudier et comprendre. Par exemple, analyser un match de football passe entres autres par déterminer le nombre de buts, le nombre de passes, la distance parcourue par les joueurs, et bien d'autres informations possibles. Autant de mesures concernant le match en lui-même, qui permettent de l'appréhender et l'étudier. L'idée à avoir en tête dans ce cas, est que même si nous n'avions pas visionné l'intégralité du match en question, nous aurions avec ces données une perception suffisante du match pour le comprendre. Dans un autre contexte, celui de la gestion d'une épidémie comme le Covid19, le nombre de personnes testées positives quotidiennement, le nombre d'hospitalisations, le nombre de décès, le facteur de contagiosité du virus, ... sont parmi les indicateurs qui nous permettent de nous formaliser en nous-même une représentation de l'évolution de l'épidémie. Dans ces cas-là (et cela est tout aussi vrai en général), adopter une approche statistique consiste à obtenir des informations concernant le phénomène en question, les modéliser mathématiquement, les étudier, puis en déduire une interprétation.

Avoir à disposition ces données, et être capable de les analyser afin de comprendre le phénomène/système considéré, voire de prédire une évolution potentielle de ce système, correspond à un enjeu actuel majeur dans de nombreux secteurs. A notre époque, cet enjeu devient de plus en plus dominant, au point que des services deviennent gratuits, de part l'intérêt d'obtenir des données concernant les utilisateurs du service. En particulier, concernant les réseaux sociaux qui sont mis à notre disposition gratuitement, il convient de garder en tête la maxime suivante :

Si c'est gratuit, c'est que vous êtes le produit !

qui indique que pour une entreprise qui gère un réseau social, le fait de mettre à disposition de tous un réseau couteux et de le maintenir, n'est rentable que dès lors qu'il peut obtenir et conserver des informations des utilisateurs. Loin de vouloir sous-entendre qu'il faille se méfier à outrance de ces réseaux, cet exemple est juste une illustration marquante de l'intérêt d'obtenir des données et des informations concernant des individus (intérêt qu'il convient d'avoir à l'esprit en tant qu'utilisateur). L'information, le renseignement (*intelligence* en anglais), correspond à une arme à posséder et à utiliser correctement. Pour cela, la statistique (collecte et analyse de la données) et l'informatique décisionnelle (automatisation des analyses et génération de rendus pertinents) sont les outils performants pour analyser les données et pour aider à la prise de décision optimale.

Quoi qu'il en soit, il existe des domaines et des questions pour lesquels l'usage de la statistique n'est pas forcément nécessaire. Par exemple, en astronomie pour décrire le mouvement des planètes, ou en mécanique newtonienne pour décrire la chute d'un corps, des équations de la physique classique permettent de décrire parfaitement (ou quasiment) le phénomène considéré. Pour autant, dans de nombreux cas, dès lors que nous ne disposons pas d'équations simples (ou de représentations mentales) qui semblent régir le phénomène, il devient alors pertinent dans ces contextes d'utiliser la statistique. Un cas illustratif est celui du débit d'une rivière en plusieurs points. Supposons connaître le débit d'une rivière donnée en un point donné (en amont). Malgré cette connaissance, il n'est pas évident d'en déduire directement quel sera le débit en un point donné en aval. On pourrait penser qu'il soit possible de facilement déduire ce débit en aval en écrivant une équation faisant intervenir le débit en amont et les facteurs en lien avec l'augmentation ou la diminution du débit de la rivière. Cependant lorsqu'on dénombre ces facteurs (infiltration dans la terre, distance entre les deux points considérés, évaporation de l'eau en fonction de la température extérieure, la topologie de la rivière autour des points considérés, la présence de rochers, la présence de pluie ou non, présences d'affluents,...), et qu'on tente de déterminer leurs impacts respectifs sur la variation du débit, on fait face à un problème beaucoup trop complexe. Une solution plus réaliste à mettre en œuvre consiste à mesurer le débit en aval et le débit en amont, puis de déterminer une liaison statistique entre ces deux grandeurs numériques. Il faut donc garder en tête, qu'il y a effectivement des contextes pour lesquels on peut se passer de l'analyse statistique au profit d'une analyse plus complète du phénomène. Quoi qu'il en soit, il arrive assez rapidement des exemples pour lesquels l'approche statistique est très efficace et beaucoup plus simple à mettre en œuvre.

Pour finir avec ce qu'il faut savoir concernant la statistique et de son utilité, formulons le assez généralement ainsi : dès qu'il est nécessaire de comprendre un phénomène complexe, une approche statistique est pertinente. Dans ce qui suit, afin de détailler plus finement le cadre de cette approche statistique, la section suivant indique qu'il est souvent nécessaire de passer par plusieurs étapes : la collecte de données, l'analyse de ces données en vue de modéliser le phénomène, et la prédiction du comportement du phénomène pour répondre à une problématique donnée.

1.2 La démarche scientifique

Afin de continuer à contextualiser le cadre de la statistique, et en particulier de la statistique descriptive, une présentation globale d'une approche statistique est donnée ci-desous étape par étape.

1. La première étape consiste à concevoir une problématique à étudier par rapport à un phénomène/système d'intérêt. L'établissement de cette problématique peut, suivant les contextes, être un travail en soi et n'être pas du tout trivial. Par exemple, supposons vouloir étudier la qualité du sommeil d'une population d'individus. Pour aller plus loin dans l'analyse, il faudrait déjà déterminer comment il serait possible d'évaluer la qualité du sommeil.
2. Il est ensuite nécessaire de déterminer quelles sont les données qui sont à collecter pour informer le décideur concernant la problématique, puis de mettre en pratique la collecte. Même si cela paraît être simple, il y a des cas pour lesquels, il n'est pas évident de déterminer quoi mesurer. En prenant de nouveau l'exemple de la qualité du sommeil, on peut se rendre compte qu'il y a tout un éventail de grandeurs physiologiques, comportementales ou environnementales à mesurer. Mais lesquelles choisir ?
3. La réflexion et la mise en pratique concernant le stockage de ces données est une étape pour laquelle il faut prendre en compte certaines contraintes. Il faut les stocker dans un format adapté, avec une structure qui rend compte des différents aspects des données, et il faudra potentiellement effectuer un pré-traitement ou un nettoyage des données. De plus, dans le cas particulier de données sensibles (voir RGPD), il convient de prendre des précautions concernant l'accessibilité à ces données. L'importance de cette étape est cruciale puisqu'une analyse correcte ne peut émaner que d'une base de données propres et bien préparées.
4. Une pré-analyse des données permet d'obtenir une représentation simple et parlante des données. Cela a pour objectif d'obtenir une première compréhension du phénomène étudié et potentiellement d'en dégager de nouvelles problématiques ou en tout cas de guider les choix qui sont à faire dans les étapes suivantes. Pour cela, il est nécessaire de déterminer des caractéristiques propres du jeu de données, comme les variations des données, la précision des mesures obtenues, ... Ne pas réaliser une pré-analyse, c'est s'exposer à passer à côté d'une information particulière dans la base de données, qui pourrait même rendre obsolète les conclusions de l'analyse finale. Par exemple, en reprenant le cas de la qualité du sommeil, supposons qu'on fasse l'impasse sur une pré-analyse qui identifierait qu'une quantité non-négligeable des personnes, dont on a mesuré le sommeil, sont sous un traitement qui est connu pour avoir un impact sur le sommeil. Sans savoir cela, l'analyse statistique considèrera à tort que l'ensemble des individus sont a priori égaux face au sommeil, ce qui pourra brouiller les résultats obtenus. A l'opposé, il pourra être pertinent d'étudier isolément chacune de ces deux sous-populations (personnes sous traitement et personnes n'ayant pas de traitement) de sorte à pouvoir correctement identifier les facteurs en lien avec la qualité du sommeil.
5. La modélisation du phénomène, au regard de la problématique et d'une compréhension experte des données, consiste à construire un modèle mathématique permettant de pouvoir apporter une réponse à la problématique posée.
6. Une phase complémentaire plus complexe est parfois nécessaire, et qui consiste à valider l'approche choisie concernant la modélisation. Pour cela des études approfondies sont à réaliser pour évaluer la robustesse de la modélisation à des variations sur les données ou sur les paramètres de la modélisation.
7. La dernière étape, à réaliser conjointement avec le commanditaire de l'étude, consiste à rendre compte des résultats, à présenter les différentes conclusions, d'en déduire les différentes décisions potentielles à prendre et d'en évaluer leurs coûts/impacts.

Dans cette procédure, la statistique descriptive correspond à la pré-analyse, la $4^{\text{ème}}$ étape. Maîtriser les outils introduits dans cette ressource pédagogique est d'intérêt pour participer correctement à ce type d'études avec une démarche scientifique.

1.3 La parabole du procès

Dans cette section, un exposé plus concret est présenté pour aider à la compréhension de la méthode statistique, mais aussi pour mettre en lumière un aspect important du contexte standard de l'application de cette méthode, et cet aspect est l'incertitude. Le contexte concret qui peut servir de parallèle est l'exemple d'un procès judiciaire. A noter qu'il n'y a pas une correspondance parfaite entre un procès et une analyse statistique, mais l'énoncé des points communs permet d'apporter un angle différent et éclairant concernant certains aspects complexes à appréhender.

Pour commencer, voici une liste des éléments importants qu'on peut retrouver dans un contexte d'une expérience et d'une analyse statistique :

Objectifs Prendre une décision optimale ou statuer sur une question, ce qui se ramène souvent à minimiser des pertes ou minimiser un risque.

Procédure *a*) Mesurer le phénomène, *b*) étudier les données collectées pour *c*) définir une généralisation/extrapolation mathématique du phénomène, puis *d*) chercher à confirmer ou à infirmer les hypothèses pré-établies, et *e*) évaluer la robustesse de l'approche utilisée.

Incertitude concernant le phénomène Lorsqu'on appréhende un phénomène, même en le percevant ou en le mesurant, nous n'avons accès qu'à une version réduite du phénomène en question. Cela vient du fait qu'on ne dispose pas d'informations concernant le phénomène dans son intégralité. Il est plutôt commun de n'avoir qu'une vision partielle du phénomène, ou autrement dit de n'avoir des données qui ne renseigne qu'une partie du phénomène.

Incertitude concernant les données Les données qui sont collectées sont communément entâchées d'une incertitude puisque ce qui nous permet d'obtenir la donnée (capteur, appareil de mesure, comptage effectué par une personne, ...) admet un niveau de précision donné (et potentiellement variable) et n'est donc pas infaillible. Autrement dit, lorsqu'on dispose d'une donnée, il est souvent nécessaire de considérer que la réalité que la donnée est censée mesurée, correspond à une variation inconnue autour de la valeur de la mesure obtenue.

Pour ce qui est du procès, supposons qu'il s'agisse d'un contexte pour lequel une personne est accusé d'avoir commis un meutre.

Objectifs Statuer sur la culpabilité de la personne accusée, en prenant garde 1) de ne pas la disculper si la personne est coupable et 2) de ne pas l'inculper si la personne est en réalité innocente.

Procédure *a*) Récolter des preuves (témoignages ou preuves physiques) en *b*) étudiant leur fiabilité afin de *c*) déterminer le profil de l'accusé et de la vraisemblance/probabilité de sa culpabilité, de sorte à *d*) faire la démonstration qu'il est ou non coupable et *e*) de pouvoir définir une décision concernant son cas en prenant en compte la cohérence de cette décision avec des procès similaires.

Incertitude concernant le phénomène Concernant la culpabilité de l'accusé, même au terme du procès, en disposant de l'ensemble des preuves, voire même de l'aveu de l'accusé lui-même, on ne pourra pas avoir une certitude absolue concernant la réalité de sa culpabilité ou non. Bien que dans certains cas, la certitude puisse être assez grande, de nombreux cas sont beaucoup plus incertains. On peut considérer qu'on ne sait jamais réellement ce qu'il en est, mais que seule notre (in)certitude varie concernant sa culpabilité, et que au-delà ou en deça d'un seuil de certitude, on pourra humblement accepter une des deux conclusions.

Incertitude concernant les données Les preuves apportées pendant le procès peuvent être assimilées à des données collectées pour renseigner sur la nature du phénomène étudié. Ces preuves ne viennent pas apporter un regard absolue sur la culpabilité de l'accusé. Par exemple, la parole d'une personne ne peut pas être reçue sans incertitude (mensonge, intérêt, mauvais souvenir, souvenirs artificiels, perception dégradée de la situation, ...). De plus, même une preuve matérielle (séquence vidéo, présence d'ADN sur la scène du crime, localisation GPS du téléphone portable de l'accusé sur la scène du crime, ...) ne peut être pris comme une preuve absolue de culpabilité. De nombreux cas pour lesquels les preuves semblaient indiquer une irréfutable culpabilité de l'accusé, qui en réalité était innocent, rendent impossible une certitude absolue.

Cette mise en mirroir, entre analyse statistique et procès judiciaire, permet de mettre en avant que dans un contexte standard de l'analyse statistique, on doit se contenter de devoir étudier des données qui sont doublement entâchées d'une incertitude qu'on ne contrôle bien souvent pas. En statistique, comme pour le cas d'un procès, ces incertitudes rendent complexes l'étude des cas et la prise de décision. Dans ces deux cas, il convient de maîtriser les notions et les outils existants pour s'assurer de fournir la meilleure analyse possible afin d'aboutir à une prise de décision optimale. Il est à noter que dans ce cas, étant donné les incertitudes inconnues et incontrôlables, l'optimalité de la décision se trouve ne même pas être évidente à définir, ni à assurer qu'elle soit atteinte.

1.4 La spécificité de la statistique descriptive

La procédure de l'analyse statistique se décompose en plusieurs étapes, parmi lesquelles une étape de pré-analyse consiste à utiliser les notions de statistique descriptive. Pour cette étape, l'intérêt est d'extraire une connaissance succincte des données, ce qui de prime à bord n'apparaît pas directement en regardant les données. Par exemple, la figure 1.1 donne un exemple de la différence qu'il y a entre un tableau de données brut et un résumé graphique des données. Pour cet exemple fictif, considérons avoir collecté pour une centaine d'étudiants, leurs choix de plats pour chaque repas dans un restaurant universitaire pendant une dizaine de jours. Pour les gérants de

	Repas_1	Repas_2	Repas_3	Repas_4	Repas_5	Repas_6
Etudiant_1	Pâtes	Pizza	Pâtes	Pizza	Pizza	Pizza
Etudiant_2	Pâtes	Pâtes	Pâtes	Pâtes	Pâtes	Pizza
Etudiant_3	Pizza	Pâtes	Pizza	Pizza	Burger	Pizza
Etudiant_4	Burger	Purée	Pâtes	Pizza	Burger	Pizza
Etudiant_5	Pâtes	Pizza	Pâtes	Pâtes	Burger	Pizza
Etudiant_6	Burger	Pâtes	Pizza	Purée	Pizza	Burger
Etudiant_7	Burger	Pâtes	Pizza	Pâtes	Pâtes	Pâtes
Etudiant_8	Pizza	Pizza	Pâtes	Poulet	Poulet	Poulet
Etudiant_9	Pizza	Burger	Pâtes	Pâtes	Pâtes	Pizza
Etudiant_10	Pâtes	Pizza	Pizza	Pâtes	Burger	Pizza
Etudiant_11	Pâtes	Burger	Burger	Pâtes	Purée	Burger
Etudiant_12	Pâtes	Pizza	Pâtes	Pâtes	Pâtes	Pizza
Etudiant_13	Pizza	Pâtes	Poulet	Poulet	Pizza	Pâtes
Etudiant_14	Pâtes	Pâtes	Burger	Pizza	Poulet	Pizza
Etudiant_15	Poulet	Pâtes	Poulet	Pâtes	Pizza	Pizza
Etudiant_16	Pizza	Pâtes	Pâtes	Pizza	Pizza	Pâtes
Etudiant_17	Pizza	Pizza	Pizza	Pâtes	Pizza	Pizza
Etudiant_18	Purée	Pâtes	Burger	Pâtes	Pâtes	Pizza
Etudiant_19	Pâtes	Pizza	Burger	Burger	Pizza	Pâtes
Etudiant_20	Poulet	Pizza	Pâtes	Pizza	Poulet	Pizza
Etudiant_21	Burger	Purée	Pâtes	Pizza	Pâtes	Poulet
Etudiant_22	Pâtes	Pizza	Pâtes	Pâtes	Pizza	Pizza
Etudiant_23	Pizza	Pizza	Pâtes	Pâtes	Pizza	Pâtes

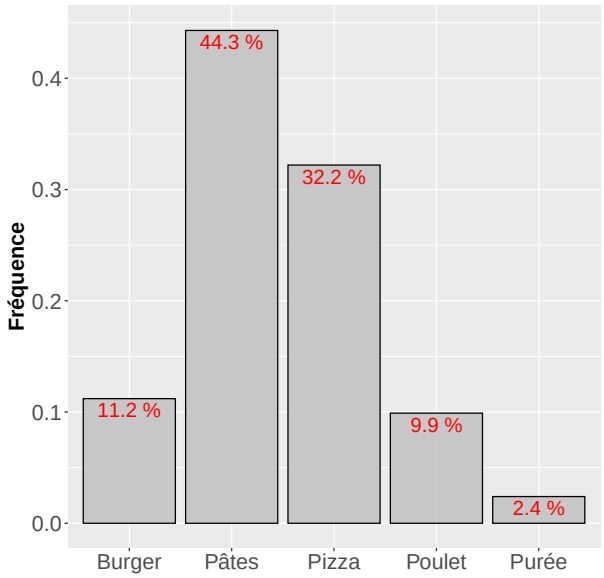


Figure 1.1 – *Sur la gauche, le tableau de données indique les plats choisis par les étudiants sur plusieurs repas dans un restaurant universitaire. Le graphique de droite est un graphique en barres permettant de visualiser rapidement la distribution des repas.*

ce restaurant universitaire, la problématique pourrait être de déterminer quels sont les choix standards des étudiants de sorte à prévoir correctement des stocks pour les semaines à venir. Dans ce cas, le tableau de données contient des centaines de lignes et des dizaines de colonnes, ce qui représente une masse d'informations imposante. Notre perception de ce tableau ne nous permet pas de visualiser ces données dans leur ensemble, et d'en obtenir rapidement une compréhension globale. Au contraire, le graphique de droite de la figure 1.1 donne des informations visuelles avec quelques chiffres clés, ce qui est plus intelligible pour un esprit humain. Le constat est simple : avoir à disposition ce tableau de données brut n'est pas du tout informatif, contrairement à un graphique qui indique les fréquences de chacunes des modalités de repas.

Pour compléter le constat précédent, et sans rentrer dans le détails des différentes étapes qui constitue l'approche statistique, un rôle potentiel d'une analyse descriptive (utilisation des outils de statistique descriptive) est, en apportant de l'information interprétable, de faire émerger de nouvelles problématiques concernant le phénomène étudié. En effet, en savoir plus concernant le contenu global des données, peut donner une perspective différente de l'étude. Par exemple, au vu du graphique de la figure 1.1, les gérants du restaurant universitaire pourraient désormais aussi vouloir clarifier la question suivante : *Est-il rentable de continuer à proposer de la purée aux étudiants ?* Pour répondre à cette problématique, il faudrait compléter l'analyse en collectant des informations concernant le prix de fabrication de la purée et le bénéfice obtenu suite à la vente d'une portion de purée.

Concernant cette approche descriptive, un aspect important à relever est qu'elle consiste principalement dans le fait de réduire l'information présente dans les données. Pour l'exemple du restaurant universitaire, on a réduit l'information que contenait le tableau de données, puisqu'en présentant les pourcentages de chacun des plats, on a gommé les informations potentielles suivantes :

- Il y a peut-être une disparité dans l'ensemble des étudiants, en terme de préférence de plats. On pourrait imaginer qu'il y a plusieurs groupes d'étudiants, avec un groupe qui ne mange que des pâtes, un groupe qui mange un peu de tout, et un groupe qui alterne entre burger et pizza. Le fait de "résumer" la situation à un pourcentage de choix pour l'ensemble des étudiants, ne permet pas d'appréhender ces potentielle différences entre des groupes d'étudiants. On rate donc ainsi, l'étude de cet aspect de la situation.
- Il pourrait aussi y avoir une différence entre les différents jours successifs (en colonnes). On pourrait très bien imaginer qu'il puisse y avoir des jours pour lesquels la pizza proposée ne correspond pas au goût des étudiants. Autrement dit, il n'y a pas une équivalence évidente entre le fait de prendre un plat au jour 1 ou au jour 2. Une analyse de l'évolution quotidienne des choix pourrait mettre en lumière des facteurs qui sous-tendent les choix des plats par les étudiants.

Ayant ainsi conscience qu'en présentant un tel graphique en figure 1.1, on perd une partie de l'information présente dans les données, il faut être convaincu que c'est à ce prix-là qu'on obtient des résultats graphiques faciles et directement interprétables. Autrement dit, il y a un compromis entre la complexité de ce qui est présenté, et notre capacité à l'interpréter. Il n'y a pas de solution optimale pour trouver un bon compromis dans ce cas-là, et la complexité de l'analyse descriptive consiste justement à trouver le juste milieu entre 1) réduire et perdre de l'in-

formation et 2) mettre à disposition un résultat interprétable. Dans le cas, d'une analyse descriptive à conduire pour un commanditaire qui doit prendre une décision concernant une problématique donnée, le statisticien doit trouver un équilibre dans ce compromis en prenant en compte :

- les indicateurs à mettre en lumière pour éclairer le décideur concernant la problématique considérée,
- les différents moyens pour représenter graphiquement ou numériquement ces indicateurs le plus simplement possible,
- les capacités d'interprétation et de compréhension du décideur, et
- une attention particulière concernant la possibilité de ne pas perdre de l'information concernant des aspects importants des données.

Dans le cadre de cette ressource pédagogique, le premier des quatre points est éclairé par l'introduction de nouvelles notions, et les trois points restants sont des compétences à acquérir avec la mise en pratique des outils introduits dans des cas concrets (voir travaux dirigés et travaux pratiques).

Le message principal qui faut garder en tête à la lecture de cette section est que l'intérêt de la statistique descriptive est de résumer l'information des données de manière pertinente de sorte à pouvoir correctement l'interpréter. D'ailleurs, on nomme "résumés statistiques" ces indicateurs qui sont utilisés pour réduire l'information de manière intelligible.

1.5 Les objectifs de cette ressource pédagogique

Cette ressource de la formation a pour objectif que l'étudiant connaisse et maîtrise les outils de statistique descriptive univariée (une variable prise isolément) et de statistique descriptive bivariée (une paire de variables).

Concernant le cas univarié, à la fin de cette ressource, l'étudiant est capable d'étudier un phénomène avec une problématique donnée, de présenter les grandeurs importantes des données et de donner une représentation graphique parlante des données. Pour ce qui est du cas bivarié, à la fin de cette ressource, l'étudiant est capable d'étudier les liens possibles entre deux variables, quelque soit le type des deux variables, de quantifier cette liaison, et de la mettre en lumière graphiquement.

Ces accomplissements sont à atteindre pour obtenir la compétence d'analyser statistiquement les données, en mettant en œuvre une analyse descriptive. Pour cela, les différentes étapes-clés de l'apprentissage à maîtriser consciemment sont :

- réaliser que les sources de données ont des caractéristiques propres à considérer,
- comprendre qu'une analyse correcte ne peut émaner que de données propres et préparées,
- comprendre l'intérêt des synthèses numériques et graphiques pour décrire une variable statistique, et
- comprendre l'intérêt des synthèses numériques et graphiques pour mettre en évidence des liaisons entre variables.

Pour illustrer dans les grandes lignes ces outils, ces objectifs et la mise en pratique de ces compétences, la section suivante propose une analyse descriptive d'un jeu de données concret.

1.6 Une application concrète : Football et Covid19

Pour finir avec cette introduction, cette section consiste en une application d'une approche statistique pour répondre à une problématique concrète. Plus précisément, avec cet exemple une grande partie des étapes détaillées en section 1.2 sont mises en œuvre, jusqu'à l'étape de l'analyse descriptive. Cette section permet ainsi d'illustrer comment se concrétise cette démarche, comment sont à utiliser les outils de statistique descriptive introduits dans ce document, et comment l'analyse descriptive s'insère, et est utile, dans une approche statistique.

1.6.1 Enoncé du contexte de l'application et de la problématique

Dans ce qui suit, le contexte étudié est celui du premier championnat français de football (Ligue 1) qui, lui aussi, a été impacté par la crise sanitaire engendrée par la pandémie de Covid19. A savoir, lors des deux saisons pendant lesquelles des restrictions sanitaires ont été prises (2019-2020 et 2020-2021), le championnat a soit du être arrêté, soit a été maintenu dans des conditions particulières. Dans ce contexte-ci, une question qui peut se poser est de savoir si certaines de ces restrictions ont eu un impact sur le déroulement "normal" du championnat. Pour le formuler autrement, aurait-on obtenu des résultats similaires si la pandémie n'avait pas eu lieu ?

Intuitivement, on peut compter parmi les contraintes qui ont pu avoir un impact sur les matchs et sur les résultats finaux des deux saisons :

- des matchs avec peu ou pas de spectateurs,
- des bulles sanitaires autour des joueurs et du staff, ou encore
- l'impossibilité de jouer pour un joueur testé positif, voire l'impossibilité de jouer pour une équipe entière dès lors qu'un certain nombre de joueurs ont été testés positifs.

Intuitivement toujours, on peut penser que le nombre de spectateurs induit une motivation particulière aux joueurs, si bien qu'en l'absence de spectateurs, la configuration d'un match n'est pas la même. A noter que les équipes qui ont les meilleurs résultats semblent globalement être celles qui sont suivies par une grande quantité de supporteurs (voir la figure 1.5, qui dans le traitement de ce contexte donne une indication claire de ce fait). Sans émettre d'hypothèses que l'un induit l'autre, ou inversément, ou que les deux s'entretiennent mutuellement, on peut se contenter de constater que les équipes du championnat sont inéquitablement suivies par une masse de supporteurs (voir la figure 1.4). Se pose alors la question de savoir pour les matchs ayant eu lieu avec peu ou pas de spectateurs, si cela a défavorisé les équipes ayant moins de supporteurs qu'habituellement.

Pour formuler plus directement la problématique qui est considérée dans cette étude et dans la suite de cette section, on peut se demander quel a été l'impact des restrictions sanitaires sur les résultats de la Ligue 1, et plus particulièrement en prenant en compte l'abscence ou la quasi-absence des spectateurs durant les matchs ?

1.6.2 La phase de collecte et de stockage

Pour tenter de répondre à cette problématique, il n'est évidemment pas possible de tester ce qu'il serait advenu, si ces contraintes n'avaient pas été mise en place (en l'absence d'une pandémie de Covid19). Pour procéder à l'analyse de l'impact de ces contraintes, il est nécessaire de faire une hypothèse à avoir en tête : on va supposer qu'il y a une tendance stable dans l'évolution du championnat, et que les saisons précédentes constituent un indicateur fiable de ce qui aurait pu se passer sans pandémie de Covid19. L'idée est donc de comparer les saisons ayant eu lieu pendant la pandémie de Covid19, avec celles ayant eu lieu juste avant.

Pour étudier cette problématique, il convient de collecter des données qui soient des quantités pertinentes et informatives concernant la problématique. Dans ce contexte, on peut penser aux grandeurs suivantes :

1. la qualité de réussite de chacun des clubs (classement, nombre de points, nombre de match gagnés ou perdus, nombre de buts mis ou encaissés),
2. les facteurs liés à l'impact des spectateurs sur l'issue du match (nombre de spectateurs, capacité du stade), et
3. les autres facteurs pouvant influer sur l'issue du match (budget des équipes, historique global en Ligue 1 pour chacune des équipes).

L'obtention des données relatives à ces grandeurs peut se faire via plusieurs moyens. Ci-dessous, une liste détaille trois possibilités dans ce contexte :

a) Le commanditaire de l'étude peut avoir au préalable l'ensemble ou une partie des données déjà collectées. Dans cet exemple, si un club est le commandidaire de l'étude, il est fort probable que le service d'analyse du club ait à disposition une grande partie des données nécessaires à cette étude.

b) La collecte peut être faite manuellement par une ou plusieurs personnes devant prendre des notes concernant le phénomène considéré. Cette possibilité produit des données qui peuvent être ponctuellement erronées. Pour cet exemple, il suffit de collecter depuis internet directement les données présentes sur des sites qui référencent les résultats du championnat de Ligue 1.

c) Suivant le contexte, les données peuvent être automatiquement acquises, soit par un capteur automatisé, par le résultat d'un algorithme de calcul, ou par un robot collecteur d'informations. Pour cet exemple, une méthode de *web scraping* peut être mise en place, ce qui consiste à développer un script qui scanne certaines pages web pour en extraire une information précise et de la mettre en forme dans une base de données.

Pour cette problématique, les données suivantes ont été collectées pour les saisons de 2014-2015 à 2020-2021 :
- le nom du club, la saison en question,
- le nombre de matchs gagnés, perdus et d'égalités pendant la saison, le nombre de buts marqués et encaissés pendant la saison, le nombre de points à la fin de la saison,
- le nombre total de spectateurs pendant la saison ayant assistés aux matchs pour chacune des équipes, la capacité d'accueil de chacun des stades,
- le budget de chacun des clubs, et des informations concernant la participation historique du club au championnat de Ligue 1 (nombre de matchs joués, nombre de points gagnés, ...)

Club	Saison	Gagne	Nul	Perdu	Buts_m	Buts_pr	Points	Nb_spectate	Budget
Lille	2020-2021	24	11	3	64	23	83	22734	147
Paris-SG	2020-2021	26	4	8	86	28	82	23306	640
Monaco	2020-2021	24	6	8	76	42	78	22060	215
Lyon	2020-2021	22	10	6	81	43	76	27610	285
Marseille	2020-2021	16	12	10	54	47	60	17196	140
Rennes	2020-2021	16	10	12	52	40	58	22386	105
Lens	2020-2021	15	12	11	55	54	57	23160	46
Montpellier	2020-2021	14	12	12	60	62	54	20502	54,5
Nice	2020-2021	15	7	16	50	53	52	10626	75

Club	Stade	Gagne_p	Nul_p	Perdu_p	Buts_marque	Buts_pris_pe	Points_perpetuel
Ajaccio	10660	131	131	224	513	738	522
Amiens	12097	25	31	48	99	144	106
Angers	17835	352	298	422	1469	1615	1354
Auxerre	23467	483	346	367	1526	1251	1795
Bastia	16480	433	301	530	1542	1813	1599
Bordeaux	42115	1069	680	757	3705	3030	3887
Brest	15931	159	165	236	617	797	642
Caen	21500	200	184	300	710	951	784
Dijon	15995	50	56	112	222	364	206

Figure 1.2 – *Les tables obtenues après collecte des données. Seules les premières lignes sont représentées.*

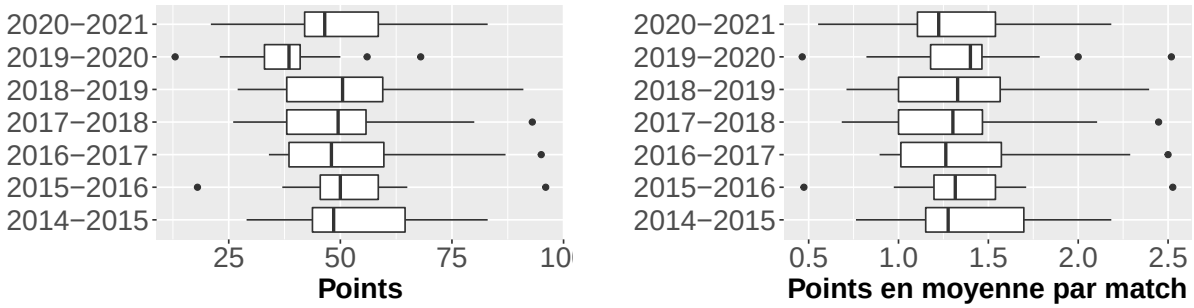


Figure 1.3 – *Le graphique de gauche correspond aux boxplots du nombre de points par saison, et celui de droite est le même graphique pour un nombre de points ramené au nombre de matchs joués durant la saison.*

1.6.3 Stockage et pré-traitement des données

La figure 1.2 montre comment les données ont été stockées en deux tables différentes. La table de gauche correspond aux informations spécifiques à chaque clubs sur chaque saison, et la table de droite contient les informations relatives aux clubs pour lequelles il n'y a pas de notions de saison. Pour que ces deux tables puissent être utilisées conjointement, il a été définie la colonne `Club` dans les deux tables, de sorte à servir d'identifiant pour croiser les tables.

Avant de pouvoir procéder à une analyse pertinente, on peut se rendre compte grâce à des représentations graphiques (outils de statistique descriptive) qu'il y a un problème dans les données. En regardant le graphique de gauche de la figure 1.3, on peut constater que pour la saison 2019-2020, la répartition des points parmi les équipes du championnat est assez inférieur aux autres saisons. Cela s'explique par le fait que durant cette saison, le championnat s'est arrété avant la fin prévue. Il en ressort donc que ces différentes saisons ne sont pas comparables en terme de points. Pour remédier à cela, une solution est de calculer pour chacune des saisons et pour chaque club, le nombre moyen de points gagnés par match, voir le graphique de droite de la figure 1.3. Cela revient à calculer une nouvelle colonne de la table de gauche de la figure 1.2 avec la formule suivante :

```
Points_moy = Points / (Gagne + Nul + Perdu)
```

où `Points_moy` est une nouvelle variable et le nom est abrégé pour "Points en moyenne". Cette étape de calcul est une étape dite de pré-traitement dans le sens où il s'agit d'appliquer un traitement au préalable de l'analyse afin de mettre en forme les données et/ou de déterminer des nouvelles grandeurs d'intérêts.

Autre exemple de ce pré-traitement, la collecte de données a été effectué sans prendre note des classements finaux de chaque saison. Dans ce cas, il est nécessaire de faire un pré-traitement (manuel ou automatique) ayant pour objectif de déterminer le classement de chacun des clubs pour chacune des saisons. D'autres indicateurs communément admis pour évaluer la chance d'un club à obtenir de bons résultats sont aussi à calculer, parmi lesquels : la différences entre le nombre de buts marqués et le nombre de but pris, ou l'augmentation du budget d'une saison sur la suivante.

Pour finir concernant cette étape, la phase de stockage des données est particulièrement importante pour réaliser correctement une analyse. De même, la phase de pré-traitement est primordiale et est généralement assez longue et fastidieuse. Il est souvent nécessaire d'effectuer un "nettoyage" des données qui consiste à détecter, puis corriger ou supprimer ou annoter, des erreurs dans la base de données. C'est aussi l'occasion d'effectuer des calculs avec les données de sorte à faire ressortir des indicateurs pertinents pour le traitement de la problématique. Le pré-traitement des données se fait généralement aussi en parallèle de l'analyse descriptive, ce qui peut mettre en lumière certains problèmes ou qui peut faire émerger le besoin de calculer certains indicateurs.

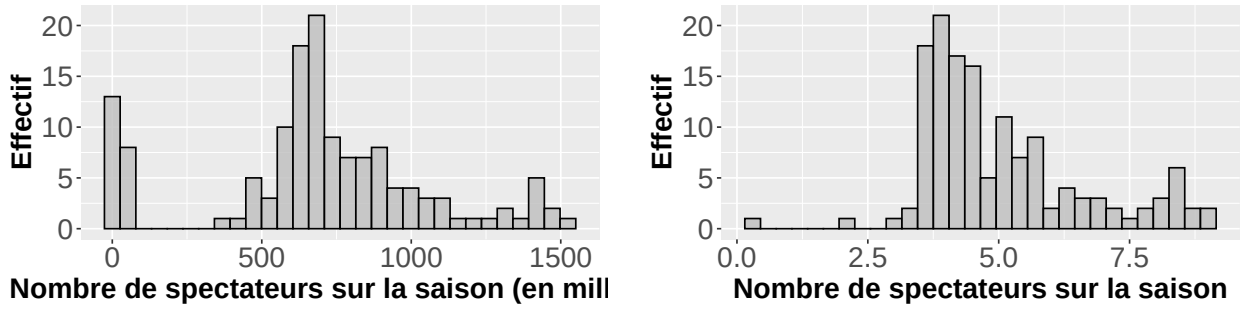


Figure 1.4 – *Histogrammes du nombre de spectateurs (en brut ou en proportion du nombre total de spectateurs), pour chaque clubs et chaque saison.*

1.6.4 Analyse descriptive

Cette étape est exploratoire et il est nécessaire d'avoir une idée a priori de quoi représenter et de comment le faire. Il n'y a pas de guide à suivre, et il s'agit de tester selon ses propres intuitions ou les recommandations du commanditaire de l'étude, avec pour objectif d'analyser les graphiques et les résultats numériques obtenus de sorte à en extraire de l'information. Pour procéder à une analyse descriptive dans le contexte de l'exemple étudié, voici les étapes qui sont suivis ci-dessous :

- représentations graphiques des quantités importantes (nombres de spectateurs, budgets, points, classement),
- représentation graphiques permettant de mettre en lumière des associations entre plusieurs variables, puis
- calculs d'indicateurs à mettre en avant pour appuyer les représentations graphiques.

Analyse descriptive univariée

Pour commencer, on représente avec les outils d'analyse descriptive univariée (voir chapitre 2) dans la figure 1.4, le nombre de spectateurs lors des saisons pré-Covid19, en y considérant le nombre brut en milliers (graphique de gauche) ou en considérant la part en pourcentage des spectateurs parmi l'ensemble des spectateurs (graphique de droite). Dans ce cas-là, l'histogramme est l'outils graphique pertinent à utiliser pour donner une visualisation intuitive de la répartition et de l'hétérogénéité des spectateurs parmi les clubs.

Pour comparer cela aux nombres de spectateurs lors de la crise sanitaire, on calcule ci-dessous les résumés statistiques pertinents : (sur la saison 2020-2021 seulement puisque les résultats de la saison 2019-2020 sont moins parlants du fait que la saison a été à moitié "normale" avant d'être arrêtée)

- les résumés de tendance centrale :
  la moyenne est de 648.87 spectateur pour la saison 2020-2021 et la médiane est 605.15,
  alors qu'avant la période "covid", la moyenne était de 18889.1 et la médiane de 18373.92,
- et les résumés de dispersion :
  pour la saison 2020-2021, l'écart-type du nombre de spectateurs est de 159.7441,
  et pour les saisons pré-Covid19, l'écart-type était de 9924.039.

Avec ces résumés statistiques, on peut constater qu'il y a effectivement une importante différence entre le nombre de spectateurs avant et pendant la pandémie de Covid19. La situation est passée d'un niveau moyen d'environ 18000 spectateurs par match, avec une grande diversité (écart-type à environ 10000) à un niveau moyen d'environ 650 spectateurs par match, avec une très faible différence entre les différents matchs. Autrement dit, les diversités passées entre les différents clubs en termes de quantités de spectateurs se sont estompées du fait des restrictions sanitaires.

Analyse descriptive bivariée

Afin de poursuivre cette analyse descriptive, on peut croiser graphiquement certaines grandeurs importantes du contexte étudié à l'aide d'outils introduits dans le chapitre 3, dédié à l'analyse descriptive bivariée. Cela permet de pouvoir représenter conjointement plusieurs variables pour déceler une liaison potentielle entre ces variables.

Il est notamment possible de constater grace à la figure 1.5 qu'il y a un lien positif entre le nombre de spectateurs et le classement final du club en temps normal. Cependant, pour la saison 2020-2021, on constate que ce lien n'existe pas. Ce graphique permet donc bien d'indiquer que la liaison pré-existante entre le nombre de spectateurs et le classement n'est plus présente. Quoi qu'il en soit, cela n'indique pas qu'il y a un lien de causalité entre ces deux grandeurs. En particulier, il se peut très probablement que la variation globale du nombre de spectateurs n'affecte

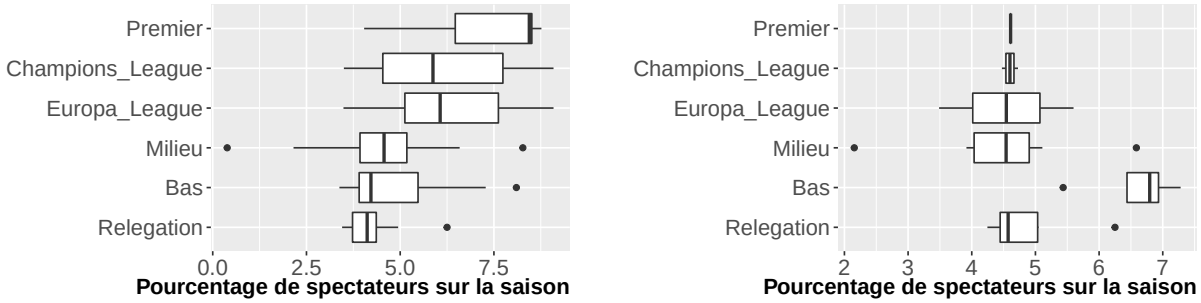


Figure 1.5 – *Représentations graphiques croisant le pourcentage de spectateurs sur la saison par rapport au classement final du club, pour la période pré-Covid19 (graphique de gauche) et pour la saison 2020-2021 (graphique de droite).*

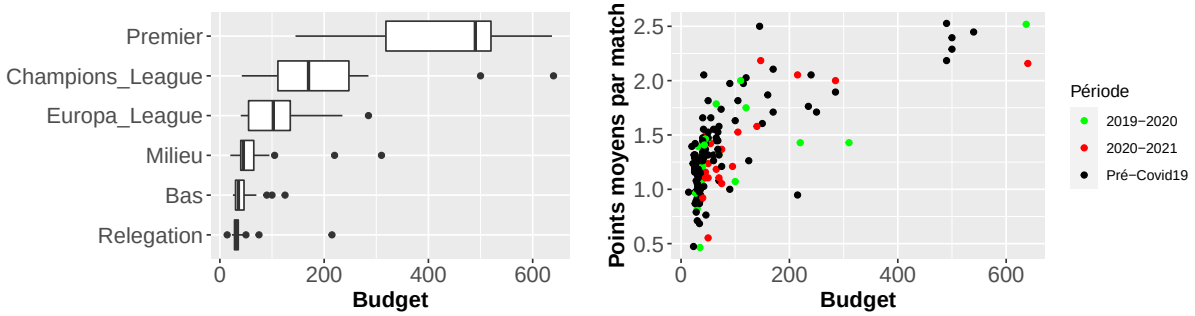


Figure 1.6 – *Représentations graphiques de l'association entre le budget des clubs et la réussite au championnat. Pour le graphique de gauche, la réussite au championnat est synthétisée en terme de positionnement dans le tableau du championnat. Pour le graphique de droite, les saisons pendant la pandémie de Covid19 sont mise en lumière.*

en rien les classements respectifs des clubs.

Pour aller plus loin dans l'analyse, on peut remarquer avec les graphiques de la figure 1.6 qu'il semble y avoir une forte association positive entre le budget du club et les résultats du club. Pour se représenter cette association, le budget est mis en regard avec deux indicateurs de réussite, la réussite au championnat sur le graphique de gauche et le nombre de points moyens par match sur le graphique de droite. On peut remarquer en particulier avec le graphique de droite, qu'il ne semble pas y avoir de différence avant la pandémie de Covid19 et pour les saisons 2019-2020 et 2020-2021, en termes d'association entre le budget et la réussite au championnat.

Renseignements obtenus et conclusions

Pour finir avec cette analyse descriptive, on peut aussi représenter les variations de classement d'une équipe d'une saison sur l'autre, au regard de la variation du budget et de la variation du nombre de spectateurs. Sur la figure 1.7, on constate qu'il ne semble pas y avoir de différence flagrante entre les deux périodes considérées. Le seul constat qui semble être à noter est qu'il semble y avoir une légère variation à la hausse du classement en présence d'une hausse du taux de variation du budget.

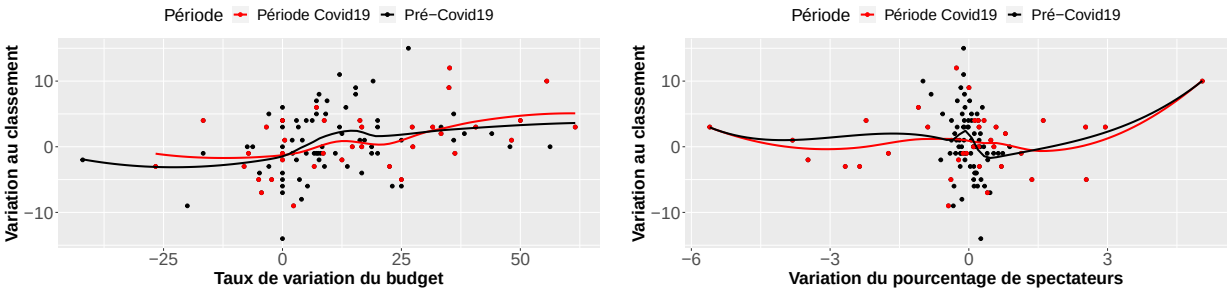


Figure 1.7 – *Représentations graphiques des variations au classement d'une saison sur l'autre en fonction de la variation du budget (graphique de gauche) et de la variation du pourcentage de spectateurs (graphique de droite).*

Pour conclure cette section, on constate gloablement qu'il y a effectivement des associations plus ou moins importantes entre le classement, le budget et la quantité de spectateurs. Pour autant, pendant la période de Covid19, la modification de la quantité de spectateurs ne semble pas avoir eu un impact sur le classement. L'hypothèse la plus vraisemblable est que, bien qu'il y ait une association statistique entre le nombre de spectateurs et la réussite du club, il ne semble pas y avoir de causalité. Tout du moins, si la quantité de spectateurs est ramenée à un niveau équivalent pour tout les clubs, cela n'induit pas de modification notable sur le classement. Avec cette analyse, il est donc déjà possible d'apporter des éléments de réponse concernant la problématique initiale, en venant dans ce cas infirmer l'intuition initiale. Cela permet donc de réorienter la problématique vers des aspects plus pertinents du contexte. Certaines des questions qui en découlent peuvent-être les suivantes :

- Quels sont les clubs qui ont échoué au classement (relativement à leur niveau habituel et à leurs investissements) et quels sont les combinaisons de facteurs qui peuvent expliquer cette contre-performance ?
- Quel devrait être l'investissement minimal pour s'assurer, avec une grande probabilité, de finir la saison dans le haut du classement ? De plus, étant donné l'évolution des clubs au fur et à mesure des saisons, quel serait la projection de cet investissement minimal pour les prochaines saisons ?
- Si pour la prochaine saison, la crise sanitaire induit que les matchs se joueront avec peu ou pas de spectateurs, pendant toute ou partie de la saison, serait-il rentable de miser sur un budget à la hausse pour cette nouvelle saison ?

Pour répondre à ce type de questions, il sera nécessaire d'utiliser des outils qui seront introduits pendant les autres semestres de la formation BUT STID.

CHAPITRE 2

Statistique univariée

Dans ce chapitre, il est question du traitement statistique unvarié sur un échantillon de données. Le terme univarié est ici à comprendre comme le fait de ne traiter que d'une (*uni-*) seule source de variabilité à la fois (*-varié*). Pour ce qui est de l'exemple du championnat, vu dans le chapitre introductif, le fait de n'étudier que la répartition des nombres de spectateurs parmi les clubs correspond à une analyse univariée. Au contraire, une analyse sera dite bivariée si elle correspond à étudier le croisement de deux facteurs (ou variables), ce qui par exemple pourra consister à évaluer si le nombre de spectateurs augmente ou diminue suivant si le budget du club est plus ou moins important (consulter le chapitre 3 pour en savoir plus). Et pour aller plus loin, une analyse melânt une multitude de facteurs est une analyse dite multivariée (au programme de $2^{\text{ème}}$ et $3^{\text{ème}}$ année).

L'objectif de la statistique descriptive univariée est de décrire, ou autrement dit de résumer, une variable à l'aide de grandeurs numériques et de représentations graphiques. Par "décrire une variable", il faut ici comprendre qu'une collection de données a été obtenue au préalable (on parle d'échantillonnage), de sorte à "mesurer" cette variable. Les indicateurs numériques et les graphiques étant calculés à partir de ces données, il est nécessaire pour la suite de connaître les notions de bases en lien avec l'échantillonnage. De la même manière il est nécessaire de savoir ce qu'est une variable et plus précisément ce qu'est une variable aléatoire.

La mise en pratique des outils introduits dans ce chapitre se fait sur deux exemples d'étude qui sont utilisés de manière récurrente. Le premier exemple est une étude de mots de passe pour en déduire les caractéristiques principales des mots de passe les plus utilisés. Le second exemple est une étude concernant les tâches solaires pour laquelle il est nécessaire de déterminer l'évolution de ces mesures au fur et à mesure des années et des centres de mesures intervenant dans l'acquisition des données.

Pour avoir une vision d'ensemble de ce chapitre, voici ci-dessous une rapide descriptif des sections suivantes. Les notions générales sont introduites en section 2.1, puis sont détaillées et complétées en section 2.2 pour le cas de données qualitatives et en section 2.3 pour le cas de données quantitatives. La section 2.4 introduit ce qu'il y a à savoir concernant les données atypiques qui ont une importance majeure dans certaines analyses. Certaines de ces notions sont mises en pratique sur des exemples en section 2.5, cette section est aussi utile pour introduire certains biais statistique importants à garder à l'esprit. Pour finir ce chapitre, la section 2.6 contient les versions "prises de note" des diapos de cours et les feuilles d'exercices qui servent de support pour les séances de travaux dirigés.

Table des matières de ce chapitre

2.1 Notions générales

Avant de pouvoir aborder le vif du sujet, à savoir les outils statistiques permettant de décrire une variable, il est nécessaire de définir ce qu'est une *variable*. Pour cela, il est nécessaire d'en passer par l'introduction de termes relatifs à une étude statistique comme la *population* et l'*échantillonnage*. En particulier, des notions assez générales en statistique sont aussi introduites (pas nécessairement en lien avec la statistique descriptive), ainsi que quelques notions de probabilités.

2.1.1 Population, échantillon et quantités importantes

La réalisation d'une étude statistique passe en premier lieu par la collecte de mesures concernant des objets d'intérêt, ou le terme objet est ici à prendre au sens large et peut englober un phénomène qu'on souhaite étudier. Dans ce cas, un objet peut être une entité matérielle (comme un être vivant ou un objet physique) ou bien être une entité immatérielle (comme une opinion).

Définition 2.1.1 (Donnée, observation ou mesure) *Perception d'un phénomène d'intérêt par un outils de mesure et donnant lieu à une valeur.*

A savoir, un outils de mesure peut être une perception humaine, un comptage, une opinion ou un capteur comme une thermomètre. La valeur qui est obtenue n'est pas forcément un nombre. Il existe plusieurs types de valeurs différents qui sont détaillés dans la section 2.1.3.

Définition 2.1.2 (Individu ou unité statistique) *Entité dont on mesure une caractéristique.*

Par défaut en statistique, on utilise le terme "individu" pour désigner une unité statistique, même s'il ne s'agit pas d'une personne ou d'un être vivant.

Exemple 2.1.3 (Fréquence de la pluie)**.** *Si on étudie la fréquence quotidienne de la pluie dans une région, pour chaque jour on doit mesurer s'il a plu ou non. Dans ce cas, une journée correspond à une unité statistique (ou un individu) et la mesure correspond à s'il a plu ou non pour cette journée.*

Il y a des contextes pour lesquels il n'est pas évident de déterminer correctement ce qu'on considère comme une unité statistique et une mesure. Il peut arriver que plusieurs points de vue soient autant valables, et la dénomination de ce qu'est l'unité statistique dépend alors de la problématique. Pour le cas de l'exemple 2.1.3, on pourrait tout aussi bien considérer que l'unité statistique est le mois et que la mesure correspond au nombre de jours pour lesquels il a plu. Considérer ce cas sous le prisme de cette vision de l'unité statistique guide l'analyse statistique vers une étude des précipitations mensuelles, et non quotidiennes. Suivant si la problématique initiale est formulée de telle sorte que l'intérêt soit porté sur la pluie au quotidien, ou sur la pluie mois par mois, l'unité statistique n'est pas la même.

A partir de la notion d'unité statistique, deux ensembles importants sont à discerner l'un de l'autre : la population et l'échantillon.

Définition 2.1.4 (Population) *Ensemble des unités statistiques.*

Définition 2.1.5 (Echantillon) *Sous-ensemble de la population constitué des unités statistiques pour lesquelles on dispose d'une mesure.*

Remarque 2.1.6. *Dans certains contextes, il est acceptable (voire utile) de considérer que la population est de taille infinie. Si elle n'est pas infinie, on note la taille de la population N. Concernant la taille de l'échantillon, on la note n.*

Exemple 2.1.7 (Fréquence de la pluie)**.** *Admettons que pour cette étude, il a été relevé pour chaque jours d'une année s'il a plu ou non. Dans ce cas-là, l'échantillon est l'ensemble des jours de cette année-là, et la population est l'ensemble de tout les jours possibles qui puissent être concernés par le phénomène de la pluie.*

Il se peut en pratique que ces deux ensembles, population et échantillon, soient confondus. Cela est possible si la population est de petite taille et que toute la population a été mesurée pour répondre au besoin de l'étude. Ce cas diffère légèrement d'une étude statistique standard, mais les mêmes outils sont à utiliser pour réaliser une analyse descriptive.

L'étude statistique a généralement pour objectif d'apporter la compréhension d'un aspect, d'une caractéristique, de la population. Cependant en pratique, on n'étudie pas réellement la population, mais plutôt l'échantillon, puisqu'on ne dispose d'information que concernant cet échantillon. L'idée est qu'on espère qu'en étudiant cet échantillon, on soit capable d'en dégager une conclusion qui puisse se généraliser à toute la population.

Exemple 2.1.8 (Fréquence de la pluie)**.** *Pour un échantillon d'une année de journées à Aurillac, on constate qu'il y a 47.7% des jours pour lesquels il a plu. A partir de ces données, et sans en savoir plus sur les années à venir et l'évolution des phénomènes climatiques, on peut envisager que pour les années suivantes, on peut s'attendre à observer autour de 172 jours de pluie.*

En passer par un échantillon est un moyen détourné de pouvoir étudier une population, puisqu'il est souvent impossible (ou trop coûteux) de mesurer toutes les unités statistiques de la population. En particulier, cela permet de pouvoir appréhender une caractéristique d'intérêt de la population.

Définition 2.1.9 (Paramètre) *Une quantité (pas forcément numérique) relative à la population.*

Dans le cadre l'exemple 2.1.8, l'étude de l'échantillon rend possible d'appréhender la probabilité qu'il pleuve dans une journée, prise au hasard, et sans informations supplémentaires concernant la journée en question ni concernant la météo des jours précédents. A noter qu'un paramètre est généralement une quantité inconnue. Non seulement, ne connaissant pas l'intégralité de la population il n'est pas possible de connaître la valeur du paramètre, mais en plus, si le paramètre était connu, il n'y aurait pas d'intérêt de réaliser une étude ayant pour objectif de le calculer.

Déduire de l'échantillon quelle pourrait être la valeur du paramètre, se ramène souvent à calculer l'expression du paramètre en question, mais en ne considérant que les valeurs qu'on a à disposition.

Définition 2.1.10 (Statistique) *Quantité calculée à partir des données de l'échantillon.*

Cette définition recouvre les résultats de tout les calculs possibles fait à partir des données. Bien que cela puisse paraître initialement étonnant, cette définition recouvre aussi l'utilisation du terme "statistique" dans la vie de tout les jours.

Exemple 2.1.11 (Match de football)**.** *Concernant une confrontation entre deux équipes de football, on désigne usuellement par statistiques, les informations suivantes : nombre de passes, nombre de buts, kilomètres parcourus, ... Si on considère cette confrontation comme une unité statistique, où la population serait alors l'ensemble des matchs possibles entre ces deux équipes, alors les mesures correspondent à nos perceptions de la confrontation. L'événement "passe" (un joueur de l'équipe A fait une passe à un autre joueur de l'équipe A) est une des mesures de la confrontation. Le nombre de passes qu'à effectuer une des deux équipes se calcule comme la somme de toutes les mesures de "passe", c'est donc une statistique.*

Parmi les statistiques calculables, certaines peuvent être utilisées à des fins descriptives (donner une résumé du match), et d'autres peuvent être utilisées pour obtenir une approximation de la valeur d'un paramètre d'intérêt (fréquence du nombre de jours de pluie pour en déduire une probabilité de pluie au quotidien).

Définition 2.1.12 (Estimateur) *Statistique dont le résultat approche la valeur d'un paramètre.*

Un estimateur peut être une quantité assez simple à calculer, comme dans les exemples 2.1.8 et 2.1.11, mais il y a aussi des contextes où les estimateurs sont complexes à construire et à calculer. Le chapitre 3 contient des estimations complexes, comme la corrélation linéaire empirique.

2.1.2 Echantillonnage et aléatoire

Apporter une réponse à une problématique, autour d'un paramètre donné, passe par calculer un estimateur. Cette procédure dépend donc de l'échantillon qui sert de base à l'analyse et bien qu'en pratique on ne dispose que d'un échantillon, il peut être pertinent de se demander si le résultat de l'analyse aurait beaucoup changé si on avait eu un autre échantillon. Une autre question importante consiste à se demander si la manière dont on a obtenu l'échantillon a un impact sur le résultat.

Définition 2.1.13 (Echantillonnage) *Méthode mise en œuvre afin d'obtenir un échantillon.*

Il existe plusieurs méthodes d'échantillonnage, dont les intérêts diffèrent suivant le contexte et la problématique. Certaines sont au programme de ressources pédagogiques en lien avec la notion d'enquête statistique, et pour cette ressource pédagogique on se contente de l'échantillonnage aléatoire simple et de d'une variante.

Définition 2.1.14 (Echantillonnage aléatoire simple) *Méthode d'échantillonnage consistant à choisir aléatoirement des individus dans une population.*

Bien que cette définition puisse paraître intuitive, elle repose sur une idée loin d'être triviale qui est qu'il est possible de faire un choix aléatoire. Pour ne pas avoir à rentrer des détails trop techniques, on suppose ici qu'il est possible d'effectuer un choix aléatoire. Le fait de considérer un choix comme "aléatoire" dans ce contexte peut être vu comme le fait que les raisons qui sous-tendent les mécanismes qui déterminent le choix n'ont aucune incidence néfaste majeure sur les résultats de l'analyse statistique.

Exemple 2.1.15 (Opinion politique)**.** *Si pour étudier l'opinion politique des français on choisit d'échantillonner des individus parmi les personnes qui nous entourent (famille, amis, ...), on obtiendra un résultat faussé, du fait que notre entourage a une forte probabilité de partager des valeurs communes, ce qui a un impact sur les opinions politiques. Dans ce cas, la méthode d'échantillonnage a un effet sur les résultats de l'analyse et on ne considère pas ici qu'il s'agit d'un échantillonnage qui a été fait sur la base de choix aléatoires.*
A l'opposé, si on sélectionne des individus en ouvrant au hasard des pages de l'annuaire, et en choisissant une ligne en fermant les yeux, les résultats de l'analyse statistique ne paraîssent pas impactés par le mécanisme de sélection (si on fait l'hypothèse qu'il s'agit d'un annuaire contenant le nom de l'ensemble des français). Il n'est pas interdit de considérer que ce mécanisme de sélection (choix de la page et de la ligne) n'est pas réellement "aléatoire", puisqu'il dépend de nos choix inconscients de s'arrêter à telle page et telle ligne. Pour autant on considérera ici le choix d'un individu de l'annuaire comme aléatoire dans le cadre de cette analyse statistique, de par le fait que les mécanismes en lien avec la sélection n'ont aucun lien avec les résultats et aucun intérêt par rapport à l'analyse statistique.

De l'utilisation de la méthode d'échantillonnage aléatoire simple, il en découle que l'échantillon obtenu est lui-même aléatoire. De même, chaque donnée de cet échantillon est considérée comme une quantité aléatoire. Par contagion, tout ce qui en découle hérite automatiquement d'une nature aléatoire.

Propriété 2.1.16 (Aléatoire) *Toute quantité qui est calculée, entièrement ou en partie, à partir de données d'un échantillon est considérée comme aléatoire.*

Il en découle qu'une statistique et un estimateur sont nécessairement des quantités aléatoires. Ce point permet de souligner un aspect important et troublant de l'analyse statistique : tout résultat obtenu avec une approche statistique est aléatoire. Ceci apporte en apparence une ombre au tableau concernant ce type d'analyse. *Comment peut-on croire une analyse qui fournit un résultat aléatoire ?* Comme il est souligné dans les ressources pédagogiques relatives aux notions de probabilités, il y a plusieurs formes d'aléatoire, et aléatoire ne signifie pas "prendre des valeurs totalement farfelues". Aléatoire au sens probabiliste (de manière complémentaire au sens statistique comme pour la propriété 2.1.16) signifie "prendre une valeur au hasard selon une répartition connue des valeurs possibles". Il est donc possible de pouvoir caractériser la forme de l'aléatoire d'une quantité aléatoire, comme celle du résultat de l'analyse statistique. Autrement dit, pour réaliser une analyse statistique correcte il faut déterminer comment se comporte l'aléatoire du résultat qui est obtenu, et cela se fait avec des outils de probabilité et des calculs mathématiques.

Exemple 2.1.17 (Opinion politique)**.** *Il est standard pour un sondage d'opinion, d'obtenir un résultat de la forme : "les intentions de vote pour le candidat A sont actuellement de* 12.78%*, avec une marge d'erreur de* 2.58%*". Cette marge d'erreur est un indicateur de la variabilité aléatoire du résultat du sondage. Une écriture plus rigoureuse du résultat, et qui permet plus directement de déceler la nature aléatoire du résultat serait : "il y a une probabilité de* 95% *que l'intervalle* $[10.2, 15.36]$ *(correspond à* 12.78 ± 2.58*) recouvre les fluctuations aléatoires du score obtenu chez les personnes sondées pour l'intention de vote pour le candidat A". On note avec cette formulation que le résultat de l'analyse (ici un intervalle) est clairement associé à une notion de probabilité et d'aléatoire.*

Pour qu'un résultat soit dûment complété avec une notion de variabilité aléatoire, la conclusion n'est pas toujours de la forme donnée dans l'exemple 2.1.17, cela dépend du contexte et de la forme de l'aléatoire en question.

2.1.3 Variable aléatoire

La section 2.1.2 introduit la notion d'aléatoire, en rapport avec l'échantillonnage. Il convient maintenant d'introduire les notions nécessaires à la mathématisation des objets aléatoires (données et statistiques).

En mathématique, il est commun d'utiliser la notation x (ou y) pour désigner une "variable". S'il s'agit ici d'une notion de variable, c'est que l'expression "$x \in [0,1]$" indique que l'objet x est définit comme étant une valeur qu'on ne connait pas, mais qui pourrait être n'importe quelle valeur entre 0 et 1. L'objet en lui-même n'est pas variable (au sens qu'il peut changer spontanément), mais dans une équation, il s'agit du terme qu'on peut choisir parmi toutes les valeurs possibles qui lui seront délimitées, de sorte à couvrir l'ensemble des résultats possibles de l'équation. Par exemple, l'expression $y = x + 1$ est une équation qui permet de définir les points d'une droite, à savoir l'ensemble $D = \{(x, y) \,|\, y = x + 1, \forall x \in \mathbb{R}\}$. Dans ce cas-là, x et y sont des variables, ce sont des symboles qui n'ont pas une valeur fixée mais qui permettent d'appréhender simultanément tout les points de la droite D.

Dans le cadre de la statistique, les variables qui sont utilisées ont pour point commun avec les variables mathématiques d'être des symboles qui représentent des objets indéterminés, mais diffèrent par le fait que les valeurs qu'elles peuvent prendre sont aléatoires.

Définition 2.1.18 (Variable aléatoire) *Une variable aléatoire X symbolise le résultat d'un phénomène aléatoire.*

Notation 2.1.19 (Variable aléatoire d'une mesure)**.** *Par convention en statistique, on note X la variable aléatoire symbolisant le résultat de l'action consistant à mesurer une unité statistique échantillonnée. De plus X_i symbolise le résultat de l'action de mesurer la $i^{ème}$ unité statistique échantillonnée.*

Notation 2.1.20 (Donnée issue d'une mesure)**.** *On utilise communément la notation x_i pour la valeur obtenue après la mesure (X_i) de la $i^{ème}$ unité statistique échantillonnée.*

Il s'agit ici d'une différence fondamentale entre la variable aléatoire et la donnée, qui justifie des notations différentes, et qui explique qu'elles interviennent différemment dans les calculs.

Exemple 2.1.21. *(Mot de passe) Supposons disposé d'un simulateur de mot de passe aléatoire. Si on évalue une première fois le simulateur on obtient le mot de passe "xkzenf". Lors d'une seconde évaluation, le mot de passe obtenu est "spvihaqz". Pour chacune des deux évaluations, il serait faux d'écrire : "X = xkzenf", ou "X = spvihaqz. Les expressions correctes sont seulement les suivantes :*

- *"X = résultat du simulateur de mot de passe",*
- *"x_1 = xkzenf" et*
- *"x_2 = spvihaqz".*

Pour distinguer l'une et l'autre des deux notions (variable aléatoire et mesure), il peut être instructif de les évaluer à l'aide de l'opérateur de probabilité.

Définition 2.1.22 (Opérateur de probabilité) *L'opérateur de probabilité $\mathbb{P}$ est une fonction qui renvoie la probabilité d'une expression (ou plutôt d'un événement).*

Remarque 2.1.23 (Notation de l'opérateur de probabilité)**.** *Il est commun de voir à la fois la notation $\mathbb{P}$ et la notation P pour désigner l'opérateur de probabilité.*

Exemple 2.1.24 (Dé équilibré)**.** *On réalise l'expérience consistante à lancer un dé équilibré à six faces et à relever le numéro inscrit sur la face du haut une fois que le dé est stabilisé. On note X le résultat du dé et après deux lancers, on obtient $x_1 = 5$ et $x_2 = 3$. Voici comment on peut correctement utiliser l'opérateur de probabilité : $\mathbb{P}(X = 4) = 1/6$. Cette expression se lit : la probabilité que le résultat de la variable aléatoire X "lancer un dé" soit 4 est de $1/6$. Pour le comprendre autrement, on peut se représenter la situation de cette expérience de la manière suivante :*

$$X = \begin{cases} 1 & \text{avec une probabilité de } 1/6 \\ 2 & \text{avec une probabilité de } 1/6 \\ 3 & \text{avec une probabilité de } 1/6 \\ 4 & \text{avec une probabilité de } 1/6 \\ 5 & \text{avec une probabilité de } 1/6 \\ 6 & \text{avec une probabilité de } 1/6 \end{cases} \tag{2.1}$$

Dans l'equation (2.1), la variable aléatoire X se définit comme pouvant prendre n'importe laquelle des six valeurs possibles avec des niveaux de probabilités donnés. Si le dé n'était pas équilibré, les probabilités dans la définition de X ne seraient pas les mêmes. L'expression $\mathbb{P}(X = 4)$ correspond à renvoyer le niveau de probabilité associé à la potentielle valeur 4. Autrement formulé, pour un évenement $X = 4$, la fonction $\mathbb{P}(.)$ renvoie la probabilité associée à la potentielle valeur 4. Pour compléter ce qu'il y a à savoir pour l'opérateur $\mathbb{P}(.)$ dans ce cas là, voici quelques expressions correctes :

- *$\mathbb{P}(X = 1) = 1/6$, ceci correspond à ce qui est expliqué ci-dessus.*
- *$\mathbb{P}(X_1 = 1) = 1/6$ et $\mathbb{P}(X_2 = 1) = 1/6$, quelque soit l'ordre de lancer de dé, les probabilités des résultats possibles restent inchangées.*
- *$\mathbb{P}(x_1 = 5) = 1$ et $\mathbb{P}(x_1 = 2) = 0$, comme x_1 indique le résultat d'une expérience qui a été réalisé, le*

résultat est connu et fixé. Il n'est donc pas probable d'obtenir une valeur différente pour x_1, qui prend la valeur 5 comme indiqué ci-dessus. Bien que la notion de probabilité et l'opérateur de probabilité ne soient pas réellement adaptés à ce type d'expression "$x_1 = 5$", on peut tout de même obtenir en toute cohérence que cette expression est de probabilité 1 et que l'expression "$x_1 = 2$" est de probabilité 0 puisque ce n'est pas possible que x_1 soit différent de 5.

2.1.4 Représentativité et généralisation

Pour compléter la section 2.1.2 et ce qu'il y a à savoir concernant les conclusions d'une analyse statistique, il est important d'introduire les notions de *généralisation* et de *représentativité*. L'analyse statistique ayant pour objectif de déterminer une valeur (approchée) d'un paramètre (inconnu car relatif à la population), à partir d'informations concernant un échantillon, cela suppose que ce qui est observé sur l'échantillon est transposable à la population.

Propriété 2.1.25 (Généralisable) *Une conclusion généralisable permet de déduire que ce qui a été constaté sur un sous-ensemble de la population, peut s'appliquer à la population toute entière.*

Toute conclusion n'est pas généalisable, et cela ne dépend pas de la conclusion en elle-même, mais principalement du résultat de la méthode d'échantillonnage mise en œuvre. Pour s'en convaincre, l'exemple 2.1.15 donne les deux cas. Le premier paragraphe décrit l'obtention d'un échantillon qui mène à une conclusion non-généralisable et le second paragraphe l'inverse. Comme cet exemple le montre indirectement, la caractéristique qui sous-tend la générabilité est la "ressemblance" qu'il y a entre l'échantillon et la population. Pourtant, il est improbable d'attendre à ce qu'un échantillon ressemble à la population, selon tout les critères possibles, et cela n'est pas non plus nécessaire pour la généralisation, cela doit dépendre de la problématique étudiée.

Propriété 2.1.26 (Représentatif) *Un échantillon est dit représentatif (de la population) pour une caractéristique donnée, si la répartition dans l'échantillon de cette caractéristique (ainsi que celles des autres facteurs en lien avec cette caractéristique) est similaire à celle de la population.*

Il n'est pas évident en pratique de vérifier cette propriété puisqu'elle dépend de la connaissance concernant la population, ainsi que de la connaissance de facteurs en lien avec la caractéristique étudiée. Cependant, des informations a priori concernant le contexte peuvent aider à évaluer la représentativité d'un échantillon.

Exemple 2.1.27 (Opinion politique)**.** *Un échantillon est collecté de sorte à appréhender la répartition de l'adhésion de principe aux partis politiques. L'échantillon obtenu se trouve ne comporter que des personnes de moins de 60 ans. Or l'âge est facteur connu comme étant en lien avec l'orientation politique, et donc cet échantillon n'est pas représentatif de la population concernant cette problématique. Pour avoir un échantillon représentatif dans ce cas-là, il faudrait au moins obtenir un échantillon pour lequel la répartition des âges dans l'échantillon, soit similaire avec la courbe des âges de la population (ce qui est connue par des études démographiques). Pour s'assurer d'être le plus proche possible de la représentativité, il faudrait de la même manière* 1) *déterminer les principaux facteurs en lien avec l'opinion politique et* 2) *s'assurer que l'échantillon et la population sont indiscernables en termes de proportions par rapport à ces facteurs.*

Afin de s'assurer l'obtention d'un échantillon représentatif dans un contexte complexe, il convient d'employer une méthode plus sophistiquée que l'échantillonnage aléatoire simple, à savoir l'échantillonnage par strate (ou échantillonnage stratifié).

Définition 2.1.28 (Strate) *Une strate est une partie de la population dont toutes les unités statistiques disposent d'un même caractère.*

Propriété 2.1.29 (Strate) *Pour un caractère donné, l'ensemble des strates sont :*
- *homogènes, au sein d'une strate les individus sont identiques par rapport au caractère considéré,*
- *distinctes, un individu ne peut pas faire partie de deux strates différentes, et*
- *supplémentaires, l'ensemble des strates recouvre l'ensemble de la population.*

Définition 2.1.30 (Echantillonnage stratifié) *Consiste à employer la méthode d'échantillonnage aléatoire simple au sein de chaque strate de la population, en échantillonnant des unités statistiques au sein de chaque strate au regard de la proportion de la strate dans la population.*

Exemple 2.1.31 (Opinion politique)**.** *Considérons vouloir échantillonner 1000 personnes dans les strates délimitées par le genre des individus (femme ou homme). En supposant que la population est constituée de moitié de femmes et de moitié d'hommes, il faut échantillonner 500 personnes parmi la sous-population des femmes et 500 personnes parmi la sous-population des hommes.*

Proposition 2.1.32 (Echantillon stratifié représentatif) *Réaliser une échantillonnage stratifié pour les principaux caractères en lien avec le paramètre d'intérêt permet d'obtenir un échantillon représentatif.*

Proposition 2.1.33 (Généralisation) *Utiliser un échantillon représentatif dans une analyse statistique est une condition nécessaire à l'établissement d'une conclusion généralisable.*

La plupart des considérations de cette section ne sont pas vérifiables en pratique. Pour autant, garder à l'esprit quels sont les rouages d'une analyse statistique qui peuvent garantir d'avoir des résultats ayant de bonnes propriétés (ici la généralisation), permet d'avoir de bonnes habitudes d'analyse et de rester vigilant quant au différents biais possibles qui ne sont pas toujours évidents à déceler.

Exemple 2.1.34 (Pandémie de Covid19)**.** *Le site* `santepublicfrance.fr` *indique que le taux de positivité des tests réalisés en France est de* 3.4%*. Une erreur de compréhension possible de ce résultat consiste à affirmer qu'il y a eu environ* 3.4% *de la population française qui a été récemment contaminée. Or, les tests n'étant réalisés que pour les personnes ayant des symptômes de la maladie, ou étant déclarées cas contact, cela constitue une sous-population ayant une probabilité plus élevée d'être malade que l'autre partie de la population. De plus, l'intensité de réalisation des tests dans cette sous-population évolue dans le temps en fonction de plusieurs critères comme la disponibilité des tests, l'intensité de la pandémie ou les décisions gouvernementales. Il est donc erroné de se baser sur cet indicateur pour affirmer quoi que ce soit concernant la dynamique de contamination de la pandémie.*

Mise en pratique des notions de la section 2.1

Exercice 2.1.1 (Insectes et déforestation)**.** Une étude a pour objet de déterminer un potentiel lien entre le niveau de déforestation d'un département et la proportion d'insectes ayant des mutations les privant de la capacité de voler. Pour cela, la proportion d'insectes ne pouvant pas voler a été relevé dans 37 localisations, ainsi que le niveau de déforestation. Dans ce contexte, quel est la population, l'échantillon et un individu ? En quoi consiste la mesure d'un individu ?

Exercice 2.1.2 (Mot de passe)**.** On souhaite étudier la fréquence des caractères pour les mots de passe. Pour faire simple, on souhaite commencer par déterminer la probabilité d'utilisation de la lettre `a` (au moins une fois) dans un mot de passe. Pour cela, on dispose d'une banque de mots de passe. Dans ce contexte, déterminez la population, l'échantillon, l'unité statistique, le paramètre et indiquez selon vous ce que serez un estimateur dans ce contexte.

Exercice 2.1.3 (Simple et stratifié)**.** Ennoncez une problématique pour laquelle un échantillonnage aléatoire simple est suffisant, et une pour laquelle il est nécessaire d'employer un échantillonnage stratifié.

Exercice 2.1.4 (Aboiements)**.** On s'intéresse à la propension à aboyer des chiens en fonction de leurs races. Quelle méthode d'échantillonnage est à utiliser ?

Exercice 2.1.5 (Dangerosité d'un carrefour)**.** On souhaite étudier les raisons pour lesquelles un carrefour peut être dangereux. Pour cela, une méthode d'échantillonnage stratifié doit être mise en place. Quelles sont les strates à considérer ?

Exercice 2.1.6 (Aléatoire ?)**.** Dans le cadre d'une analyse statistique, qu'est-ce qui peut être considéré comme aléatoire parmi les éléments suivants :

- la population,
- l'échantillon,
- une individu de la population,
- un individu de l'échantillon,
- la mesure d'un individu de la population,
- la mesure d'un individu de l'échantillon,
- un estimateur,
- un paramètre,
- la conclusion de l'analyse,
- la problématique.

Exercice 2.1.7 (Gains des streamers)**.** Une étude s'intéresse aux gains des streamers francophones et à l'importance relatives des différentes sources de revenus. Pour cela, l'étude considère les gains et les sources de revenus des 100 streamers les plus populaires. Deux conclusions sont données par l'étude en question, et pour chacune d'elles indiquez si elles sont généralisables à l'ensemble des streamers francophones :

1. L'analyse permet de déterminer que la source de revenus la plus importante est la publicité.
2. Au regard des données obtenues, la rémunération brute d'un streamer n'excède pas `xxx` dollars par mois.

2.2 Description de données qualitatives

Cette section est dédiée à l'analyse descriptive de données qualitatives. Les données qualitatives englobent plusieurs types de données qui sont détaillées par les définitions 2.2.1, 2.2.2 et 2.2.4.

Définition 2.2.1 (Donnée binaire) *Mesure associée à un phénomène pouvant prendre uniquement deux états.*

Si la mesure d'un phénomène renvoie des valeurs suivantes parmi les valeurs suivantes, il s'agit de données binaires : 0 ou 1, vrai ou faux, échec ou succès, ...

Définition 2.2.2 (Donnée qualitative ou catégorielle) *Mesure associée à un phénomène pouvant prendre un nombre positif et limité d'états.*

Voici des exemples de données catégorielles : la première lettre d'un mot de passe, le candidat pour qui un individu vote, ...

Définition 2.2.3 (Modalité ou niveau) *Valeur possible prise par une donnée catégorielle.*

Définition 2.2.4 (Donnée ordinale) *Donnée catégorielle pour laquelle il y a un ordre naturel entre les différentes modalités.*

La mesure de satisfaction d'un produit ou d'un service dont les modalités sont "pas du tout satisfait", "as satisfait", "neutre", "satisfait", "très satisfait" est une donnée ordinale. Il y a un ordre de préférence entre les différentes modalités.

Bien que ces types de données diffèrent légèrement par nature, elles peuvent toutes être englobées sous la terminologie de données catégorielles et leurs traitements est similaire dans le cadre d'une analyse descriptive.

2.2.1 Indicateurs numériques

Pour résumer l'information importante concernant la modalité d'une variable qualitative, il faut principalement décrire l'intensité d'utilisation de cette modalité par les individus de l'échantillon.

Définition 2.2.5 (Effectif d'une modalité) *Le nombre d'individus de l'échantillon possédant la modalité en question.*

Définition 2.2.6 (Fréquence d'une modalité) *Le proportion d'individus de l'échantillon possédant la modalité en question.*

Exemple 2.2.7 (Mot de passe)**.** *Dans une banque de mots de passe contenant* 1000 *mots de passe, 537 d'entre eux contiennent au moins une fois la lettre* `a`*. Dans ce cas, 537 est l'effectif de la modalité "*`a`*" et* 53.7% *est sa fréquence.*

Remarque 2.2.8. *On peut aussi parler de fréquence absolue à la place d'effectifs et alors de fréquence relative à la place de fréquence.*

Ces indicateurs sont utiles pour renseigner des informations concernant des quantités (théoriques) concernant la population, autrement dit par définition : des paramètres .

Définition 2.2.9 (Probabilité d'une modalité) *Correspond à la fréquence des unités statistiques dans la population qui possèdent cette modalité.*

Exemple 2.2.10 (Pandémie de Covid19)**.** *En prenant un échantillon de la population française, on peut déterminer la fréquence des personnes contaminées dans cet échantillon. Si l'échantillon est représentatif par rapport à la propension des individus à être contaminés, on pourrait en conclure que la fréquence observée renseigne efficacement sur la fréquence de contamination parmi toute la population française. On ne connait pas cette fréquence dans la population et dans ce contexte, la probabilité d'être malade pour une personne prise au hasard dans cette population se définit comme cette "fréquence dans la population". Autrement dit, le paramètre "probabilité d'être malade" peut être estimé par la statistique "fréquence d'individus contaminés dans l'échantillon".*

Définition 2.2.11 (Distribution d'une variable aléatoire qualitative) *Correspond à l'ensemble des probabilités de chacune des modalités de la variable aléatoire qualitative.*

La distribution renseigne sur tout ce qu'il y a à savoir concernant l'aspect aléatoire d'une variable qualitative. Cette

distribution est inconnue puisqu'elle dépend de la population, mais on peut calculer la distribution empirique.

Définition 2.2.12 (Distribution empirique d'une variable aléatoire qualitative) *Correspond à l'ensemble des fréquences de chacune des modalités dans l'échantillon de la variable aléatoire qualitative.*

Remarque 2.2.13 (Empirique et théorique)**.** *Ces termes sont utilisés dans le cadre de cette ressource pédagogique de sorte à distinguer les deux cas suivants :*

- *empirique : fait référence à toutes les quantités qui dépendent des observations ou de l'expérience,*
- *théorique : fait référence à tout ce qui est inconnu et qui dépend de la population.*

Les quantités théoriques sont généralement les objectifs d'une problématique et les quantités empiriques sont celles obtenues par l'analyse statistique pour estimer les quantités théoriques. Plus précisément, ces quantités théoriques (paramètres) sont des valeurs qui caractérisent la variable aléatoire associée au phénomène mesuré. Autrement dit, calculer les quantités empiriques c'est s'approcher des quantités théoriques inconnues, et donc approcher une quantité qui donne une information importante et caractéristique du phénomène étudié.

Remarque 2.2.14 (Statistique et probabilités)**.** *Les notions théoriques, comme la définition 2.2.9 et la définition 2.2.11, sont des notions de probabilités, qui sont donc aussi au programme d'une ressource pédagogique dédiée aux probabilités. Ces notions sont aussi vues dans le cadre de cette ressource pédagogique afin d'expliciter le lien qu'il y a entre les notions de statistiques (valeurs empiriques) et ces notions de probabilités (valeurs théoriques).*

Pour compléter la notion de distribution empirique, on utilise le mode pour désigner la modalité la plus représentée.

Définition 2.2.15 (Mode) *Correspond à la modalité ayant la fréquence la plus élevée.*

Exemple 2.2.16 (Mot de passe)**.** *Parmi les mots de passe de la banque de mots de passe (n = 339391), on s'intéresse aux lettres plutôt qu'à l'ensemble des caractères possibles et la table 2.1 donne les effectifs des lettres utilisées dans ces mots de passes, ainsi que leurs fréquences parmi toutes les lettres utilisées. En particulier, la colonne "Fréquence" renseignent les fréquences de toutes les modalités possibles, et il s'agit donc de la distribution empirique de la variable aléatoire "lettre d'un mot de passe". Autrement dit, cela indique avec quelle importance se répartissent les lettres de l'alphabet dans les mots de passes. On trouve ici que le mode de la distribution est la lettre* `e`*, comme attendu pour des mots de la langue française.*

Caractère	Fréquence	Effectif	Caractère	Fréquence	Effectif
a	9.6 %	290861	n	6.74 %	204281
b	1.84 %	55769	o	7.68 %	232893
c	4.44 %	134635	p	3.26 %	98907
d	3.04 %	92038	q	0.17 %	5038
e	10.28 %	311650	r	7.26 %	220127
f	1.17 %	35339	s	6.31 %	191218
g	2.16 %	65576	t	6.8 %	206044
h	2.49 %	75570	u	3.53 %	106918
i	9.11 %	276180	v	1.22 %	37084
j	0.21 %	6375	w	0.58 %	17700
k	0.74 %	22417	x	0.29 %	8723
l	5.39 %	163462	y	1.94 %	58814
m	3.22 %	97589	z	0.52 %	15887

Table 2.1 – *Distribution des lettres de l'alphabet parmi les mots de passe de la base de données.*

2.2.2 Indicateurs graphiques

Les résumés graphiques possibles pour une variable aléatoire qualitative consistent à donner une représentation visuelle de la distribution empirique. Plusieurs types de graphiques sont possibles pour représenter une même distribution empirique. Le choix de quel graphique utilisé dépend du contexte et les illustrations de la figure 2.1 permettent de donner un aperçu des raisons qui guident ce choix.

Le principal intérêt est de sélectionner un graphique de sorte à ce que l'information importante de l'échantillon apparaissent de manière évidente. Un graphique bien choisi et bien réalisé apporte une plus-value non-négligeable à une étude statistique. Prendre du temps pour trouver la bonne représentation graphique peut apparaître comme une perte de temps, lors de l'analyse. Cependant, un mauvais graphique peut valoir des retours néfastes à l'équipe en charge de l'analyse statistique. Voici ci-dessous un guide pour choisir parmi les graphiques possibles :

Diagramme en barre A éviter si les fréquences sont similaires parmi les différentes modalités.

Diagramme en barre en coordonnées polaires Utile s'il s'agit d'une variable ordinale et s'il y a plus d'une dizaine de modalités. Par exemple si les différentes modalités correspondent aux 12 mois de l'année.

Diagramme empilé A éviter s'il y a plus d'une dizaine de modalités. Utile s'il s'agit d'une variable ordinale.

Diagramme en camembert A éviter s'il y a plus d'une dizaine de modalités et s'il y a des modalités ayant des fréquences très proches de 0. Exception à cela s'il n'y a que 3 ou 4 modalités et que l'une d'entre elle admet une très faible fréquence ou une fréquence très élevée.

Mise en pratique des notions de la section 2.2

Exercice 2.2.1 (Type de données)**.** Parmi les énoncés suivants, dites de quel type de données qualitatives il s'agit :

- Mention au bac.
- Lancer d'une pièce.
- Rôle dans une équipe de football.
- Réussite ou non à un examen.
- Genre d'un film.

Exercice 2.2.2 (Effectif et fréquence)**.** Pour les données suivantes, calculez les effectifs et les fréquences de chacune des modalités. De plus, dites quel est le mode.

Film	Genre
La mémoire dans la peau	Thriller
Le Dernier des Mohicans	Western
Fight Club	Action
Taxi	Action
Le Petit Monde de don Camillo	Comédie
Matrix	Science-fiction
Near death experience	Drame
Le Corniaud	Comédie

Film	Genre
Et pour quelques dollars de plus	Western
Le Dernier pour la route	Drame
Sixième Sens	Thriller
Le Transporteur	Action
Je suis heureux que ma mère soit vivante	Drame
Clean	Drame
La Grande Vadrouille	Comédie
Il était une fois dans l'Ouest	Western

Exercice 2.2.3 (Distribution empirique)**.** Tracer un diagramme en barre des données de l'exercice 2.2.2.

Exercice 2.2.4 (Probabilité d'un échantillon)**.** On considère une variable aléatoire qualitative X ayant K modalités et dont la distribution théorique est noté $(p_1, p_2, \dots, p_K)$, où $p_k = \mathbb{P}(X = k)$, pour $k = 1, \dots, K$.

1. Calculez la probabilité d'obtenir la modalité 1. Même chose pour la modalité k.
2. Calculez la probabilité d'obtenir une modalité entre 1 et k.
3. Que vaut $\sum_{k=1}^{K} p_k$?
4. Pour deux tirages indépendants de la variable aléatoire X, calculez la probabilité d'observer deux fois la même modalité.
5. Calculez la probabilité d'obtenir la modalité 1 puis la modalité 2.
6. Calculez la probabilité d'obtenir la modalité 1 et la modalité 2, indépendamment de l'ordre de tirage.
7. Calculez la probabilité d'observer un échantillon donné de taille n.

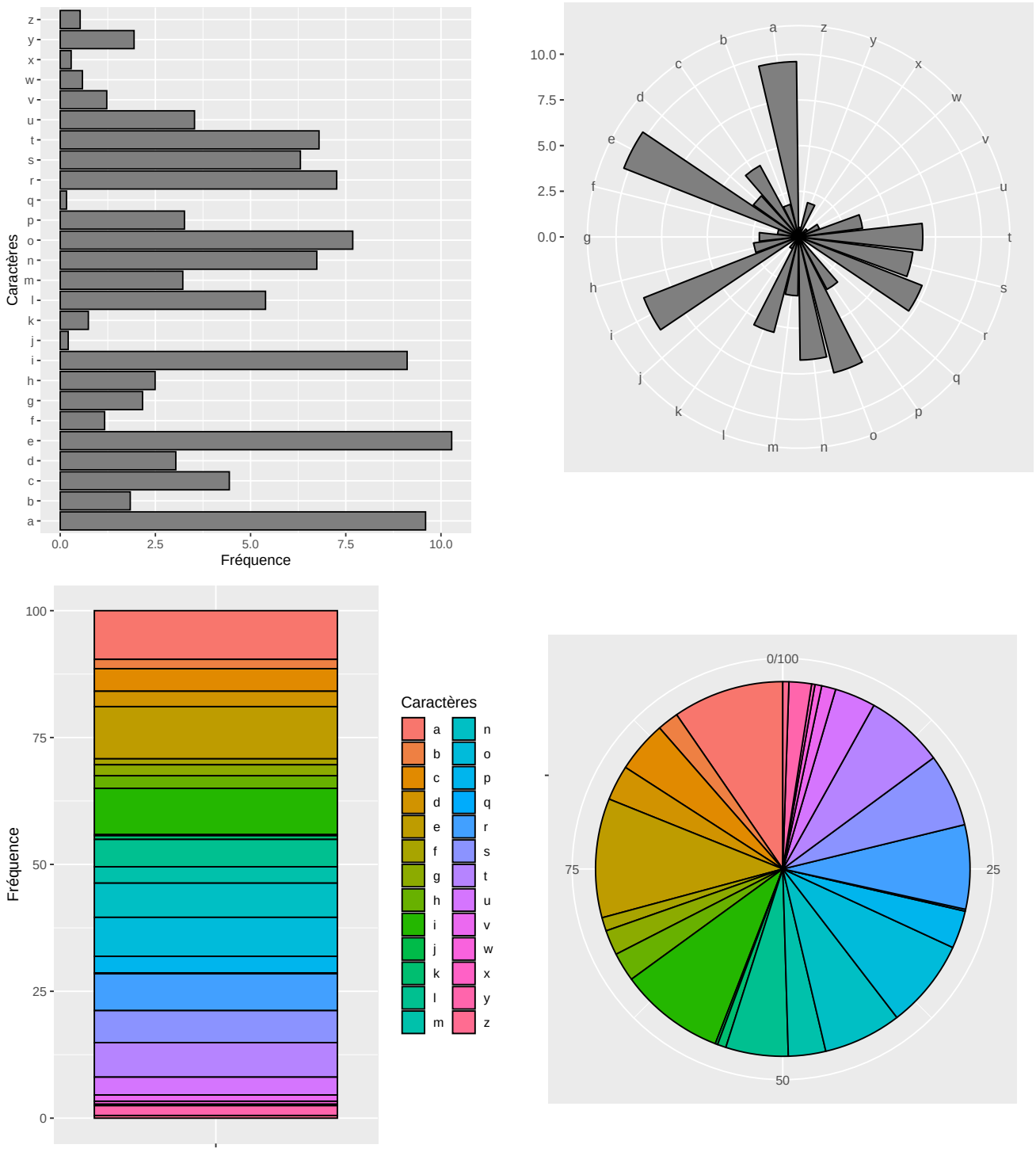


Figure 2.1 – *Les différentes outils graphiques pour représenter la distribution empirique d'une variable aléatoire qualitative. Le graphique en haut à gauche est un diagramme en barre, celui en haut à droite est le même diagramme mais en coordonnées polaires, le graphique en bas à gauche est un diagramme empilé et le celui en bas à droite est un diagramme en camembert (pie chart en anglais).*

2.3 Description de données quantitatives

Par le terme "données quantitatives", cela désigne toutes les mesures numériques. Dans cette famille de données, et il faut distinguer quelques unes d'entre elles pour la suite.

Définition 2.3.1 (Donnée quantitative discrète) *Mesure d'un phénomène donnant lieu à des valeurs numériques entières.*

Cette catégorie regroupe principalement les données issues d'un comptage (nombre d'étudiants dans une promotion par exemple). Certaines données peuvent faire penser à des données quantitatives discrètes alors qu'elles ne le sont pas. Par exemple, la position d'une équipe dans un classement n'est pas une donnée quantitative discrète. Pour constater une différence, il suffit de se demander si cela aurait un sens d'effectuer des additions entre ces données, si cela a du sens il s'agit de données quantitatives discrètes, si cela n'en a pas il s'agit d'un autre type de données. Pour l'exemple de la position dans un championnat, additionner la première position et la deuxième position n'a pas de sens, et il s'agit dans ce cas d'une donnée ordinale (et donc qualitative).

Remarque 2.3.2 (Discret)**.** *L'éthymologie du mot "discret" fait référence à une notion de discontinu, de séparé et de distinct. Il faut donc ici comprendre qu'une donnée discrète correspond à mesurer des éléments distincts.*

Définition 2.3.3 (Donnée quantitative continue) *Mesure d'un phénomène donnant lieu à des valeurs numériques réelles.*

On peut distinguer ce type de données des données quantitatives discrètes en se demandant s'il est possible d'obtenir une mesure avec des décimales. Si c'est le cas il s'agit d'une donnée quantitative continue.

Exemple 2.3.4 (Température)**.** *On dispose de la collection de données suivante :* $15, 24, 29, 31, 28$*. Ces données correspondent à des mesures de températures maximales sur plusieurs jours consécutifs. Bien qu'on n'ait observé que des valeurs entières, il n'est pas impossible d'observer la valeur* 25.4*, il s'agit donc de données quantitatives continues.*

Exemple 2.3.5 (Nombre de pattes)**.** *On dispose des données suivantes qui correspondent aux nombres de pattes d'un animal :* $2, 2, 4, 8, 2$*. Comme cela n'a pas de sens d'avoir* 2.5 *pattes, il ne s'agit pas de données quantitatives continues.*

Le propos du reste de cette section est de déterminer ce qu'il y a à faire pour effectuer une analyse descriptive sur des données quantitatives. Pour certains aspects, il est nécessaire de faire la distinction entre "discret" et "continu", mais pour d'autres les traitements sont communs.

2.3.1 Catégorisation

Le premier traitement possible pour ce type de données est de se ramener à des données qualitatives pour appliquer les outils descriptifs de la section 2.2. Cette conversion de données quantitatives en données qualitatives est en apparance assez simple puisqu'il suffit de définir des classes de valeurs (notion similaire à celle de strate, voir la définition 2.1.28).

Exemple 2.3.6. *On peut convertir les données de l'exemple 2.3.4 en définissant les classes de valeurs suivantes :*
- *"froid", pour les valeurs inférieures à 20 degrés,*
- *"tempéré", pour les valeurs entre 20 et 30 degrés, et*
- *"chaud", pour les valeurs supérieures à 30 degrés.*

L'échantillon devient donc la collection de données catégorielles suivante : froid, tempéré, tempéré, chaud, tempéré. Il est alors possible de calculer les effectifs des modalités, la distribution des données et de recourir à des représentations graphiques comme le diagramme en barre.

Cependant, cette approche comportent deux limitations importantes. La première est dans le choix des classes. De prime abord cela peut paraître simple. Pour autant, il faut avoir à l'esprit qu'il n'y a souvent pas de choix de classes naturel. De plus, le choix des classes de valeurs peut influer lourdement sur le résultat de l'analyse descriptive. Pour être guidé dans le choix de ces classes de valeurs, voici plusieurs conseils :

Classe naturelle Si une segmentation naturelle est possible, il vaut mieux se baser sur cette segmentation pour construire les classes de valeurs.

Homogénéité Au sein d'une même classe de valeurs, les individus doivent être homogènes par rapport aux caractéristiques étudiés.

Classe rare Il faut éviter de constituer des classes de valeurs ne contenant que très peu d'individus.

Exemple 2.3.7 (Pandémie de Covid19). *On dispose de données concernant les personnes ayant contractées une forme sévère de la maladie, notamment concernant des caractéristiques importantes comme l'âge et le niveau de vaccination. Plutôt que de décrire la propension à faire une forme sévère en fonction de l'âge (de 12 à 106 ans), il est plus parlant de faire des classes d'âges. Dans ce cas, on peut constituer des classes d'âges par dizaine d'années : 10-19, 20-29, ..., 90-99 et 100-109. L'idée derrière cette catégorisation est qu'il n'y a a priori pas de différences marquantes entre une personne ayant 21 ans et une personne ayant 22 ans. Des différences marquantes devraient apparaître entre dizaines d'années environ. Il peut être pertinent de faire des classes d'âges pour manquer la différence en termes de propension à faire une forme sévère.*

Bien que cette catégorisation puisse paraître naturelle et suffisante, deux problèmes subsistent. Le premier est que parmi la classe d'âges "10-19", il y a de grandes différences en terme de vaccination puisqu'en dessous de 16 ans, le mineur ne peut pas décider seul de se faire vacciner. Comme cela a probablement un impact en terme de propension à contracter une forme sévère de la maladie, il n'est donc pas pertinent de ne faire qu'une classe d'âge, et on préfèrera la scinder en deux classes : 10-15 et 16-19.
Le second problème concerne les classes d'âges des personnes les plus âgées. Comme il n'y a que peu de personnes ayant plus de 90 ans, par rapport aux nombres de personnes constituant les autres classes, il convient de rassembler les classes "90-99" et "100-109" en une seule classe : "90+".

Exemple 2.3.8 (Congés maladie). *Le rectorat est en charge (entre autres) de relever et d'étudier la quantité de jours de congés maladie pris (sur l'année) par les contractuels de la fonction publique. Les données vont de 0 à 46 et il convient dans ce contexte de pouvoir faire des classes de valeurs pour fournir des résultats assez parlant. Cependant, il n'y a pas de classe naturelle dans ce contexte, il est alors nécessaire d'effectuer une analyse particulière (avec des données supplémentaires) pour savoir comment définir des classes homogènes et en évitant de construire des classes rares.*

Un autre aspect de la catégorisation peut être à la fois un avantage comme un inconvénient, il s'agit de la réduction de l'information (voir la section 1.4). La catégorisation induit une réduction de l'information puisque passer d'une collection de valeurs différentes à une classe, gomme les disparités entre ces différentes mesures. Pour l'exemple 2.3.6, on peut constater que passer des valeurs 24, 29 et 28 à la classe "tempéré" induit que l'analyse qui en suit, ne fait pas la différence entre une température de 24 degrés et une température de 29 degrés. Il est commun que l'utilisation de résumés statistiques induisent une réduction de l'information. Cependant la catégorisation peut se voir comme une phase de pré-traitement qui vient au préalable, et en plus, de l'utilisation de résumés statistiques. Autrement dit, cela correspond à réduire plusieurs fois l'information présente dans les données. Si la catégorisation se fait selon de mauvais choix de classes, cette perte d'information peut avoir un effet néfaste sur la qualité des résultats de l'analyse. Cependant, quand il est possible de faire des classes pertinentes (comme dans l'exemple 2.3.7), cela peut fournir une analyse descriptive parlante. Pour savoir quand il est recommandé d'utiliser une approche de catégorisation pour le traitement de données quantitatives, il faut garder à l'esprit qu'il faut le faire si seulement les critères vus en début de cette section sont bien applicables (classe naturelle ou pertinente, homogénéité et classe rare), et si les résultats obtenus par une analyse descriptive sans catégorisation seraient trop complexes et pas assez parlants. La complexité doit être pensée sous le prisme de la personne qui va consulter les résultats de l'analyse. S'il s'agit d'une personne lambda, certains aspects d'une analyse descriptive de données quantitatives sans catégorisation (voir section 2.3.2 et section 2.3.3) peuvent ne pas être intuitifs ou interprétables, et alors il est plus adapté d'en recourir à la catégorisation. Par exemple, il est fréquent que des résultats statistiques à présenter au grand public fassent intervenir une catégorisation (voir dans les médias ou pour les messages de santé publique).

Si les données quantitatives sont converties en données qualitatives, les outils de la section 2.2 sont à utiliser pour réaliser une analyse descriptive. Dans ce qui suit, les outils introduits sont ceux à utiliser s'il convient de ne pas faire de catégorisation des données quantitatives.

2.3.2 Indicateurs numériques

Comme pour le cas de données qualitatives, il y a derrière l'utilisation d'indicateurs numériques, le fait de donner une valeur approchée d'un paramètre théorique correspondant à une des caractéristiques principales de la variable aléatoire étudiée (voir la section 2.2.1 et en particulier les remarques 2.2.13 et 2.2.14). Pour introduire les indicateurs numériques utilisés pour la description des données quantitatives, il est alors nécessaire au préalable d'introduire les notions théoriques utiles pour caractériser une variable aléatoire quantitative.

2.3.2.1 Indicateurs de position et de dispersion

Puisque les notions théoriques dépendent des valeurs des individus dans la population, il est nécessaire d'introduire une notation pour ces valeurs.

Notation 2.3.9 (Valeur d'un individu de la population)**.** *Pour le $i^{ème}$ individu de la population, on note par $\tilde{x}_i$ (se dit "x-i-tilde") sa mesure pour le phénomène considéré.*

Notation 2.3.10 (Valeur possible dans la population)**.** *S'il y a p valeurs possibles dans la population, on note y_i la $i^{ème}$ de ces valeurs.*

Une première caractéristique d'une variable aléatoire quantitative est l'espérance, donnant une information sur la "position centrale de la distribution".

Définition 2.3.11 (Espérance) *L'espérance, notée μ et prononcée "mu", correspond à la moyenne des valeurs des individus de la population.*

Cette espérance peut se voir comme la valeur qu'on s'attend à obtenir si on calcule la moyenne d'un échantillon avec beaucoup de données. On dit que cela indique la "position" de la distribution puisque c'est la valeur autour de laquelle on s'attend à observer la plupart des valeurs de l'échantillon. Par exemple, si l'espérance de la taille d'un homme adulte est 1.70 mètre, et qu'on échantillon quelques hommes dans la population (des adultes) on s'attend à ce que leurs tailles se répartissent autour de 1.70 mètre. Il y en aura qui auront une taille inférieure à cette espérance et d'autres qui auront une taille supérieure.

Notation 2.3.12 (Opérateur de somme)**.** *Une somme de termes indicés $x_1 + x_2 + \cdots + x_N$ peut se réécrire sous la forme $\sum_{i=1}^{N} x_i$ ou $\sum_{i=1}^{N} x_i$, ce qui se lit "la somme des termes x_i pour i allant de 1 à N".*

L'expression de l'espérance peut donc s'écrire :

$$\mu = \frac{1}{N}(\tilde{x}_1 + \tilde{x}_2 + \cdots + \tilde{x}_N) = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \tag{2.2}$$

où N est le nombre total d'individus dans la population. Il est possible de réécrire cette espérance sous une forme plus synthétique et plus habituelle, en notant qu'il y a dans la population des individus qui devraient avoir la même mesure. On peut alors effectuer une factorisation dans l'expression 2.2. Par exemple, si on suppose que la population contient 20 individus, pour laquelle

- la première valeur possible y_1 apparait 2 fois dans la population,
- la deuxième valeur possible y_2 apparait 6 fois dans la population,
- la troisième valeur possible y_3 apparait 12 fois dans la population, et
- qu'il n'y a pas d'autres valeurs

alors l'espérance s'écrit :

$$\mu = \frac{2}{20} \times y_1 + \frac{6}{20} \times y_2 + \frac{12}{20} \times y_3$$

Les fréquences $\frac{2}{20}$, $\frac{6}{20}$ et $\frac{12}{20}$ sont des fréquences de mesures dans la population et il s'agit donc de probabilités (voir définition 2.2.9) et alors :

$$\mathbb{P}(X = y_1) = \frac{2}{20}, \quad \mathbb{P}(X = y_2) = \frac{6}{20} \text{ et } \quad \mathbb{P}(X = y_3) = \frac{12}{20}.$$

L'équation (2.2) se réécrit alors comme :

$$\mu = y_1\mathbb{P}(X = y_1) + \cdots + y_p\mathbb{P}(X = y_p) = \sum_{i=1}^{p} y_i\mathbb{P}(X = y_i) \tag{2.3}$$

où p correspond au nombre de valeurs possibles dans la population.

La notation μ est une notation de l'espérance générique pour toute variable aléatoire quantitative, mais pour parler de l'espérance d'une variable aléatoire spécifique, notée X, on utilise la notation suivante.

Définition 2.3.13 (Opérateur d'espérance) *Pour une variable aléatoire X, son espérance est donnée par l'expression $\mathbb{E}(X)$ où $\mathbb{E}$ est l'opérateur d'espérance.*

Pour les calculs du paragraphe précédent, une hypothèse importante a été effectuée et n'a pas été mentionnée directement. Il s'agit de l'hypothèse qu'il y a p valeurs possibles dans la population (voir notation 2.3.10), ce qui induit nécessairement qu'il s'agisse d'une variable aléatoire quantitative discrète. A l'opposé s'il s'agit d'une

variable aléatoire quantitative, il y a une infinité de valeurs possibles dans la population, si bien que les expressions obtenues (2.2) et (2.3) ne sont pas valides (pour des raisons qui sont à voir dans la ressource pédagogique dédiée aux probabilités). Dans ce cas, il vient que :

- le pendant de l'opérateur de somme $\sum$ est l'opérateur d'intégrale $\int$, et
- le pendant de la notion de probabilité $\mathbb{P}(X = y_1)$ est la notion de densité de probabilité $f_X(x)$.

Pour synthétiser, la propriété suivante indique l'expression de l'espérance dans chacun des deux cas.

Propriété 2.3.14 (Expression de l'espérance) *Si X est une variable aléatoire quantitative discrète, son espérance a pour expression :*

$$\mathbb{E}(X) = \sum_{i=1}^{N} y_i \mathbb{P}(X = y_i).$$

Si X est une variable aléatoire quantitative continue, son espérance a pour expression :

$$\mathbb{E}(X) = \int x f_X(x)\,\mathrm{d}x$$

Un second indicateur important d'une variable aléatoire quantitative est la variance. Cet indicateur, complémentaire de l'espérance qui indique la "position centrale" des valeurs de la population, indique à quel point il y a une dispersion dans les valeurs possibles de la population.

Définition 2.3.15 (Variance) *La variance, notée σ^2, correspond à l'espérance des (carrés des) distances entre les valeurs possibles de la population et l'espérance.*

Pour le formuler en des mots plus simples, la variance correspond grossièrement à la moyenne des distances qu'il y a entre chaque valeurs possibles et centre de ces valeurs. Si pour une variable aléatoire, la variance est plutôt faible (proche de 0), cela indique que parmi les individus de la population, il n'y a pas une grande variété dans les mesures et que toutes les mesures de la population sont très proches de la moyenne de la population (espérance). A l'opposé si la variance est grande (positive et éloignée de 0), cela correspond à une population pour laquelle il y a des individus ayant des mesures très éloignées de l'espérance.

Définition 2.3.16 (Opérateur de variance) *Pour une variable aléatoire X, sa variance est donnée par l'expression $\mathbb{V}(X)$ où $\mathbb{V}$ est l'opérateur de variance.*

Propriété 2.3.17 (Expression de la variance) *Si X est une variable aléatoire quantitative discrète, sa variance a pour expression :*

$$\mathbb{V}(X) = \sum_{i=1}^{N} \big(y_i - \mathbb{E}(X)\big)^2 \mathbb{P}(X = y_i) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big].$$

Si X est une variable aléatoire quantitative continue, son espérance a pour expression :

$$\mathbb{V}(X) = \int \big(x - \mathbb{E}(X)\big)^2 f_X(x)\,\mathrm{d}x = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big].$$

Les détails de calculs permettant de passer d'une expression à l'autre dans les formules de la propriété 2.3.17 ne sont pas à comprendre ou à connaître dans le cadre de cette ressource pédagogique.

Les précédents paramètres (espérance et variance) sont des grandeurs qui permettent de caractérister l'aléatoire d'un phénomène. En pratique ces paramètres ne sont pas connus et il convient de calculer des indicateurs empiriques pour les approcher. L'indicateur de position qui permet d'approcher l'espérance, de la variable aléatoire relative au phénomène, est la moyenne des données de l'échantillon.

Définition 2.3.18 (Moyenne) *Pour un échantillon de taille n, dont les mesures obtenues sont $x_1, x_2, \ldots, x_n$, la moyenne des données est notée $\overline{x}_n$ et son expression est :*

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Remarque 2.3.19 (Notation de la moyenne). *Suivant le contexte, s'il n'est pas nécessaire d'indiquer que la moyenne des données se calcule à partir d'une quantité de n données, il est possible d'utiliser la notation $\overline{x}$ à la place de $\overline{x}_n$.*

Remarque 2.3.20 (Moyenne et espérance). *Il est important de ne pas confondre la moyenne (quantité empirique) et l'espérance (quantité théorique) bien que ces deux notions se ressemblent.*

Exemple 2.3.21 (Tâches solaires). *Pour étudier l'évolution des tâches solaires, une permanence scientifique a relevé la quantité mensuelle de tâches solaires de 1749 à aujourd'hui. Au-delà de 1960, le centre de mesure est passé d'une ancienne installation (Swiss Federal Observatory à Zurich) à une installation plus moderne (Tokyo Astronomical Observatory). On traite ici seulement des données présentées dans la table 2.2, pour les années 1960 et 1961, en se demandeant si le passage d'un centre de mesure à un autre a induit un changement dans la qualité des mesures. La moyenne des données se calcule ici par :*

$$\overline{x}_n = \frac{1}{24}(146.30 + 106 + \cdots + 32.60 + 32.90) = 83.07917.$$

La moyenne peut aussi se calculer par année, on note $\overline{x}_{1960}$ et $\overline{x}_{1961}$ ces deux moyennes annuelles, et on obtient : $\overline{x}_{1960} = 112.275$ et $\overline{x}_{1961} = 53.88333$, ce qui semble pointer une nette différence entre ces deux années.

Moyenne mensuelle	Année	Mois	Moyenne mensuelle	Année	Mois
146.30	1960	Jan	57.90	1961	Jan
106.00	1960	Feb	46.10	1961	Feb
102.20	1960	Mar	53.00	1961	Mar
122.00	1960	Apr	61.40	1961	Apr
119.60	1960	May	51.00	1961	May
110.20	1960	Jun	77.40	1961	Jun
121.70	1960	Jul	70.20	1961	Jul
134.10	1960	Aug	55.80	1961	Aug
127.20	1960	Sep	63.60	1961	Sep
82.80	1960	Oct	37.70	1961	Oct
89.60	1960	Nov	32.60	1961	Nov
85.60	1960	Dec	39.90	1961	Dec

Table 2.2 – *Nombres moyens de tâches solaires observées par mois pendant les années 1960 et 1961.*

L'indicateur de dispersion qui permet d'approcher le paramètre de variance est la variance empirique calculée sur les données de l'échantillon.

Définition 2.3.22 (Variance empirique) *Pour un échantillon de taille n, dont les mesures obtenues sont $x_1, x_2, \ldots, x_n$, la variance empirique des données est notée s^2 et son expression est :*

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2.$$

Grâce au théorème suivant, on peut obtenir une version simplifiée de l'expression de s^2, simplifiée dans le sens où il est nécessaire de faire moins d'opération pour obtenir le résultat numérique.

Théorème 2.3.23 (Köning-Huygens) *La variance peut se réécrire :*

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

et la version empirique de cette formule donne :

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}_n^2$$

De cet indicateur, on préfère parfois en déduire un autre indicateur, l'écart-type.

Définition 2.3.24 (Ecart-type) *L'écart-type s est la racine carré de la variance :*

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2}.$$

L'écart-type a l'avantage de s'exprimer dans la même unité que celle des données, comme la moyenne, contrairement à la variance. Plus précisément, il correspond à la moyenne des distances qu'il y a entre une donnée prise au hasard et la moyenne des données. Pour certains raisons théoriques (au programmme d'autres ressources pédagogiques dédiées à la statistique), on peut utiliser une variante de la variance empirique (qui peut aussi se décliner en un écart-type corrigé).

Définition 2.3.25 (Variance empirique corrigée) *Pour un échantillon de taille n, dont les mesures obtenues sont $x_1, x_2, \ldots, x_n$, la variance empirique corrigée des données est notée $\hat{\sigma}^2$ et son expression est :*

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2.$$

Exemple 2.3.26 (Tâches solaires)**.** *Avec les mêmes données que l'exemple 2.3.21, on obtient les mesures de dispersion suivantes :*

- *variance empirique :*

$$s^2 = \frac{1}{24}\left((146.30 - 83.07917)^2 + (106 - 83.07917)^2 + \ldots (39.90 - 83.07917)^2\right) = 1115.047$$

- *variance empirique corrigée :*

$$\hat{\sigma}^2 = \frac{1}{23}\left((146.30 - 83.07917)^2 + (106 - 83.07917)^2 + \ldots (39.90 - 83.07917)^2\right) = 1163.527$$

- *écart-type :* $s = \sqrt{s^2} = 33.39231$.

*En moyenne, l'écart entre une donnée et la moyenne (*83.07917*) est d'environ 33. En calculant les variances des données pour chacune des années, qu'on note s^2_{1960} et s^2_{1961}, on obtient : $s^2_{1960} = 361.1269$ et $s^2_{1961} = 164.1731$. Il y a ici une différence notable entre ces deux années, pour l'année 1960 la quantité de tâches solaires a été plus variable et plus élevée par rapport à celle de l'année suivante. En consultant les données, on constate effectivement que pour l'année 1960, les données sont très dispersées, en allant de 80 à 150 et de 30 à 80 pour l'année 1961.*

Une dernière mesure de dispersion permet de mesurer la dispersion relativement à l'échelle des données.

Définition 2.3.27 (Coefficient de variation) *Si X est une variable aléatoire quantitative, son coefficient de variation a pour expression :*

$$c_v(X) = \frac{\sqrt{\mathbb{V}(X)}}{\mathbb{E}(X)}$$

Définition 2.3.28 (Coefficient de variation empirique) *Pour un échantillon de taille n, dont les mesures obtenues sont $x_1, x_2, \ldots, x_n$, le coefficient de variation empirique des données est noté $\hat{c}_v$ et son expression est :*

$$\hat{c}_v = \frac{s}{\overline{x}}$$

Cette mesure peut exprimer un écart à la moyenne en terme de pourcentage. A noter que ce coefficient est très sensible à des variations de mesures si la moyenne est proche de 0. De plus, ce coefficient perd du sens si les mesures étudiées n'ont pas toutes le même signe. Il est plus cohérent de l'utiliser si les données sont par définition toutes positives ou toutes négatives, comme par exemple la température en degré Kelvin.

2.3.2.2 Indicateurs relatifs à la distribution

Comme le montre l'exemple étudié dans la section 2.5.3, il se peut que les indicateurs de position (comme la moyenne) ne soit pas les plus intéressants suivant le contexte. Dans ce cas, on peut utiliser un autre indicateur pour déterminer quelle est la valeur la plus représentative.

Définition 2.3.29 (Fonction de masse) *Pour une variable aléatoire quantitative discrète X dont les valeurs possibles appartiennent à un ensemble discret K, la fonction de masse f_X donne les probabilités de chacun des éléments de l'ensemble K :*

$$f_X(k) = \mathbb{P}(X = k), \qquad \textit{pour tout } k \in K.$$

Définition 2.3.30 (Densité de probabilité) *Pour une variable aléatoire quantitative continue X dont les valeurs possibles appartiennent à l'intervalle $\mathcal{I}$, la fonction de densité de probabilité indique pour une valeur $x \in \mathcal{I}$ à quel point il est probable d'observer une valeur proche de x :*

$$f_X(x) = \lim_{h \to 0} \frac{\mathbb{P}(x \leq X \leq x + h)}{h}$$

Remarque 2.3.31 (Densité de probabilité). *Cette définition n'est pas celle utilisée pour définir correctement la notion de densité, mais elle a le mérite d'être simplifiée. De plus, l'équation donnée dans la définition n'a pas pour intérêt de donner une méthode calculatoire pour calculer la densité de probabilité d'une variable aléatoire en un point x, mais elle a seulement pour utilité de donner l'intuition qu'il s'agit de la probabilité d'observer une valeur proche de x, où la zone "proche" de x est indéfiniment petite.*

En pratique, on ne peut pas calculer les quantités théoriques des définitions 2.3.29 et 2.3.30. Il faut calculer les versions empiriques de ces quantités.

Définition 2.3.32 (Fonction de masse empirique) *Pour un échantillon de taille n, dont les mesures discrètes obtenues sont $x_1, x_2, \ldots, x_n$, la fonction de masse empirique est :*

$$\hat{f}_X(k) = \frac{n_k}{n}$$

où n_k est le nombre de mesures x_i qui valent k.

Concernant la fonction de densité de probabilité empirique, qu'on note aussi $\hat{f}_X$, des méthodes standards sont à utiliser (à voir en séance de Travaux Pratiques), mais elles font intervenir des notions qui ne sont pas au programme de cette ressource pédagogique et qui ne sont donc pas détaillées dans ce document.

Propriété 2.3.33 (Somme et intégrale) *Pour une fonction de masse f_X et une fonction de masse empirique $\hat{f}_X$ on a que :*

$$\sum_{k=1}^{N} f_X(k) = 1 \quad \text{et} \quad \sum_{k=1}^{N} \hat{f}_X(k) = 1$$

où N est le nombre de valeurs possibles. Cette propriété fait écho au fait que la somme des probabilités vaut 1. De plus, pour une fonction de densité de probabilité f_X et une fonction de densité empirique $\hat{f}_X$, on a que :

$$\int f_X(x)\,\mathrm{d}x = 1 \quad \text{et} \quad \int \hat{f}_X(x)\,\mathrm{d}x = 1.$$

Définition 2.3.34 (Mode) *Le mode d'une variable aléatoire quantitative X est la valeur qui maximise la fonction $\hat{f}_X$, que ce soit une fonction de masse empirique si X est discrète ou une fonction de densité de probabilité empirique si X est continue.*

Remarque 2.3.35 (Calcul du mode). *Si la variable X est discrète, le mode peut se calculer simplement, mais si la variable X est continue, il faut utiliser un ordinateur pour calculer une densité de probabilité empirique.*

Un indicateur supplémentaire, et moins utilisé, sert à quantifier une potentielle asymétrie dans la répartition des données. Dans la population, les valeurs possibles des mesures se répartissent autour de l'espérance, avec une intensité quantifiée par la variance. Il peut y avoir plusieurs scénarios possibles de répartition autour de l'espérance :

- il peut y avoir plus de valeurs inférieures à l'espérance (asymétrie négative),
- il peut y en avoir à peu près autant qui soient inférieures que supérieures (asymétrie nulle), et
- il peut y avoir plus de valeurs supérieures à l'espérance (asymétrie positive).

Pour mesurer l'asymétrie, la quantité théorique est le coefficient d'asymétrie.

Définition 2.3.36 (Coefficient d'asymétrie) *Le coefficient d'asymétrie d'une variable aléatoire X est :*

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}}\right)^3\right]$$

La version empirique et calculable en pratique de ce coefficient est l'indicateur suivant, qui permet de quantifier la potentielle asymétrie de la répartition des données autour de la moyenne.

Définition 2.3.37 (Coefficient d'asymétrie empirique) *Pour un échantillon de taille n, dont les mesures obtenues sont $x_1, x_2, \ldots, x_n$, le coefficient d'asymétrie empirique est :*

$$G_1 = \frac{n^2}{(n-1)(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^3}{\left(\hat{\sigma}^2\right)^{3/2}}$$

Exemple 2.3.38 (Tâches solaires). *Avec les mêmes données que l'exemple 2.3.21, on obtient le résultat suivant pour le coefficient d'asymétrie empirique : $G_1 = 0.2466619$. Ce résultat est cohérent avec le fait qu'il y a 13 mesures qui sont supérieures à la moyenne, contre 11 qui sont inférieures. Il y a donc bien une asymétrie positive.*

2.3.2.3 Indicateurs basés sur les quantiles

Il existe d'autres indicateurs numériques qui peuvent renseigner concernant la position, la dispersion, ainsi que des aspects extrêmes d'un échantillon et qui sont tous basés sur le calcul de quantiles empiriques. Quoi qu'il en soit, on peut commencer par les introduire sans définir la notion de quantile.

Définition 2.3.39 (Médiane) *Correspond à la valeur qui sépare en deux parts égales l'échantillon, et qu'on note M.*

La médiane est un indicateur de position au même titre que la moyenne. Ces deux indicateurs ne donnent pas le même résultat et ils ont des propriétés différentes. Dans certains contextes, on préfèrera utiliser l'un plutôt que l'autre, voir la section 2.5.1 pour plus de détails.

Définition 2.3.40 (Série ordonnée) *Correspond aux valeurs de l'échantillon dont les valeurs ont été rangées par ordre croissant (de la plus faible à la élevée). Par convention, on note $x_{(i)}$ la $i^{ème}$ valeur de la série ordonnée, qui n'est pas forcément égale à la $i^{ème}$ valeur de l'échantillon.*

Il y a plusieurs conventions possibles pour calculer la médiane et dans le cadre de cette ressource pédagogique, on considère qu'elle se calcule de la manière suivante.

Remarque 2.3.41 (Calcul de la médiane)**.** *Si n est impair (autrement dit il existe k tel que $n = 2k + 1$), alors il est possible de déterminer une valeur présente dans l'échantillon qui sépare l'échantillon en deux parts égales contenant k mesures chacune. Il suffit donc de calculer la série ordonnée de l'échantillon et de déterminer la $(k + 1)^{ème}$ valeur pour trouver la médiane de l'échantillon. Autrement dit, la médiane est obtenue grâce à la formule suivante*

$$M = x_{(k+1)}.$$

Si n est pair (autrement dit il existe k tel que $n = 2k$), alors il n'est pas possible de déterminer une valeur présente dans l'échantillon qui sépare en deux parts égales l'échantillon. Pour cela, on détermine la médiane par la valeur se trouve au milieu entre la $k^{ème}$ valeur et la $(k + 1)^{ème}$ valeur de la série ordonnée. La médiane se détermine donc par :

$$M = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

Définition 2.3.42 (Quartiles) *Le premier quartile, noté Q_1, est la valeur qui sépare l'échantillon en deux parts telles que 25% des données de l'échantillon soient inférieures à Q_1 et que 75% soient supérieures.*

Le troisième quartile, noté Q_3, est la valeur qui sépare l'échantillon en deux parts telles que 75% des données de l'échantillon soient inférieures à Q_3 et que 25% soient supérieures.

Remarque 2.3.43 (Quartile et médiane)**.** *Le $2^{ème}$ quartile coïncide avec la médiane.*

Les quartiles sont les indicateurs de forme de la distribution empiriques des données. En effet, suivant la position des quartiles par rapport à la médiane, on peut en déduire une potentielle asymétrie positive ou négative. Par exemple, si la distance entre Q_1 et la médiane M est plus grande que la distance entre Q_3 et M, cela indique qu'il y a une asymétrie négative. Et inversément pour avoir une asymétrie positive.

Remarque 2.3.44 (Calcul des quartiles)**.** *Comme pour la médiane, le calcul des quartiles dépend de la taille de l'échantillon n. Pour la premier quartile, on calcule $k = \frac{n+3}{4}$. Si on obtient une valeur entière, alors le résultat est la $k^{ème}$ valeur de la série ordonnée :*

$$Q_1 = x_{(k)}.$$

Si ce n'est pas le cas, on note i la valeur entière de k et il faut déterminer quelle valeur entre la $i^{ème}$ et la $(i+1)^{ème}$ valeur de la série ordonnée est le premier quartile. On obtient le résultat avec la formule suivante :

$$Q_1 = (i + 1 - k)x_{(i)} + (k - i)x_{(i+1)}$$

S'il s'agit de calculer le $3^{ème}$ quartile, la procédure est la même ave $k = \frac{3n+1}{4}$.

Définition 2.3.45 (Ecart inter-quartile) *On note* IQR *(pour "InterQuartile Range" en anglais), l'écart qu'il y a entre les deux quartiles :* $\text{IQR} = Q_3 - Q_1$.

Il s'agit d'un indicateur de dispersion qui renseigne sur l'intensité d'éloignement des données autour de la médiane.

La notion (théorique) de quantile permet de généraliser ces différents indicateurs.

Définition 2.3.46 (Quantile) *Pour une variable aléatoire quantitative X, le quantile de niveau α est la valeur q_α qui vérifie : $\mathbb{P}(X \leq q_\alpha) = \alpha$.*

Un quantile est donc une quantité qui permet de séparer la distribution d'une variable aléatoire en deux parts, de sorte à ce que la partie inférieure recouvre un total de probabilité donné.

Définition 2.3.47 (Quantile empirique) *Pour un échantillon de données, le quantile empirique de niveau α est la valeur $\hat{q}_\alpha$ qui sépare en deux parts l'échantillon de telle sorte que $\alpha\%$ des données observées soient inférieures à $\hat{q}_\alpha$ et que $(100-\alpha)\%$ soient supérieures.*

Remarque 2.3.48. *En pratique, on peut ne pas pouvoir trouver une portion de l'échantillon qui fasse exactement $\alpha\%$, ce qui arrive dès lors que $n \times \frac{\alpha}{100}$ n'est pas une valeur entière. Si c'est le cas, il faut déterminer une valeur entre deux mesures de l'échantillon comme pour le calcul des quartiles (voir les remarques 2.3.35 et 2.3.44).*

Remarque 2.3.49 (Calcul des quantiles empiriques)**.** *On calcule $k = n \times \frac{\alpha}{100}$ et on note i la valeur entière de k. Si k est un nombre entier (et donc $k = i$) alors*

$$\hat{q}_\alpha = x_{(i)}.$$

Sinon, on obtient le résultat avec la formule suivante :

$$\hat{q}_\alpha = (i + 1 - k)x_{(i)} + (k - i)x_{(i+1)}$$

Par définition, comme par construction, on peut trouver les relations suivantes :
- $Q_1 = \hat{q}_{25\%}$,
- $M = \hat{q}_{50\%}$, et
- $Q_3 = \hat{q}_{75\%}$.

Exemple 2.3.50. *Avec les mêmes données que l'exemple 2.3.21, on obtient les résultats suivants pour quelques quantiles empirique :*

α	$\hat{q}_\alpha$
10%	38.58
20%	50.02
Q_1	53.00
30%	56.22
40%	62.72
M	77.40
60%	87.20
70%	105.24
Q_3	110.20
80%	120.02
90%	125.12

2.3.3 Représentations graphiques

2.3.3.1 Boxplot

Un premier résumé graphique possible pour des données quantitatives est le boxplot (”boîte à moustache” en français), qui donne une idée de la répartition des données. La figure 2.2 donne un exemple de boxplot calculé sur les données de tâches solaires de l'exemple 2.3.21. Ce graphique comporte plusieurs délimitations qu'on détaille ci-dessous en balayant le graphique de gauche à droite :
- La première délimitation à gauche du graphique correspond à une borne minimale estimée par la formule suivante, et notée $b_{\min}$:
$$b_{\min} = \max(x_{(1)}, Q_1 - 1.5 \times \text{IQR}).$$
- La seconde délimitation correspond au premier quartile Q_1.
- La troisième délimitation correspond à la médiane M.
- La quatrième délimitation correspond au troisième quartile Q_3.
- La cinquième délimitation correspond à une borne maximale estimée par la formule suivante, et notée $b_{\max}$:
$$b_{\max} = \min(x_{(n)}, Q_3 + 1.5 \times \text{IQR}).$$

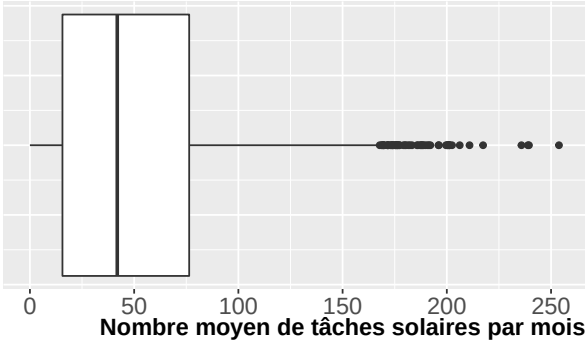


Figure 2.2 – *Boxplot des données de l'exemple 2.3.21.*

Pour rappel, $x_{(i)}$ correspond à la $i^{\text{ème}}$ valeur de la série ordonnée, et donc $x_{(1)}$ et $x_{(n)}$ correspondent respectivement aux valeurs minimale et maximale observées. De plus, les points qui apparaissent sur le graphique du boxplot sont les données observées qui dépassent des bornes estimées $b_{\min}$ et $b_{\max}$.

2.3.3.2 Histogramme

L'histogramme est un graphique dont un exemple est donné dans la figure 2.3 et qui donne une idée plus précise que le boxplot concernant la répartition des données. Pour obtenir un histogramme, il faut catégoriser les données en des classes de valeurs et représenter l'effectif (voir graphique de gauche de la figure 2.3) ou la fréquence (voir graphique de droite) de chacune des classes de valeurs, avec un diagramme en barre. On note H_k la hauteur de la $k^{\text{ème}}$ barre de l'histogramme.

- S'il s'agit d'un histogramme d'effectif, on a $H_k = n_k$, où n_k est le nombre de mesures appartenant à la $k^{\text{ème}}$ classe de valeurs.
- S'il s'agit d'un histogramme de fréquence, on a $H_k = \frac{f_k}{\ell_k}$ où f_k est la fréquence $k^{\text{ème}}$ classe de valeurs et ℓ_k est la largeur de la $k^{\text{ème}}$ barre, de sorte à ce que l'aire de la $k^{\text{ème}}$ barre soit égale à la la fréquence f_k et que la somme des aires des barres fasse 1.

Pour calibrer la largeur des classes de valeurs (et donc des barres de l'histogramme), il y a des méthodes automatiques, mais il est possible de les choisir soi-même. Cependant, ce choix impacte la lisibilité du graphique, la figure 2.4 donne deux exemples de mauvais choix, un avec trop peu de classes et l'autre avec trop.

Exemple 2.3.51. *Avec les données des tâches solaires, la première classe déterminée automatiquement est la classe* $[-4.5, 4.5]$. *Il y a 287 mesures dans cette classe de valeurs, et c'est bien la hauteur qu'on retrouve pour la première barre de l'histogramme d'effectif. La fréquence de cette classe est* $f_k = 287/3168 = 0.0915404$ *et la hauteur* H_1 *de l'histogramme de fréquence est donc* $H_1 = 0.0915404/9 = 0.01041667$, *ce qu'on retrouve bien sur le graphique de droite de la figure 2.3.*

2.3.3.3 Densité empirique

Un autre outils de représentation graphique consiste à représenter une estimation de la densité de probabilité f_X. Pour cela des méthodes complexes sont disponibles et faciles d'utilisation, mais leur principe n'est pas détaillé dans ce document. La figure 2.5 donne un exemple de ce qu'on peut obtenir avec cette approche. Le graphique de droite de cette figure est une superposition de la densité de probabilité empirique et de l'histogramme des fréquences. Avec ce graphique, on peut avoir l'intuition du lien suivant : la densité empirique peut se voir comme la courbe limite lorsque les largeurs des barres ℓ_k d'un histogramme tendent vers 0.

Mise en pratique des notions de la section 2.3

Exercice 2.3.1 (Tâches solaires)**.** Effectuez une catégorisation des données de tâches solaires, de l'exemple 2.3.21, avec des classes de valeurs de longueurs 20 : $[30, 50[, [50, 70[, \ldots$ Représentez les fréquences des classes de valeurs avec un diagramme en barre.

Exercice 2.3.2 (Suisse)**.** Calculez les résumés statistiques numériques pour chacune des deux variables "Fertilité" et "Agriculture" des données suivantes correspondant à des mesures socio-économiques de provinces Suisse en 1888 :

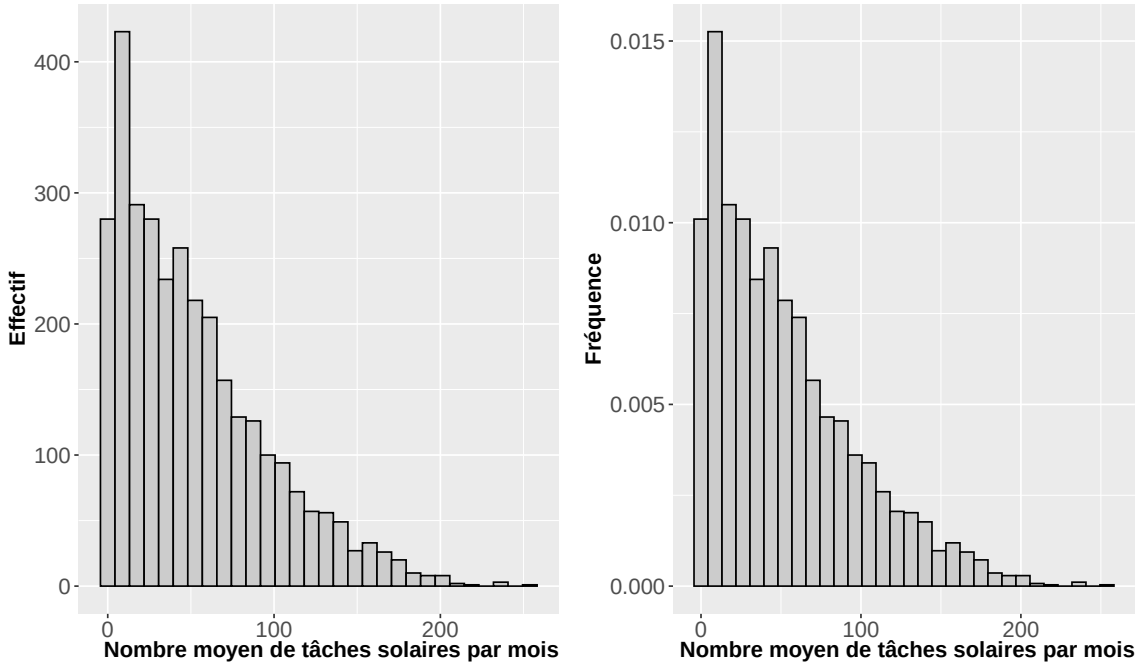


Figure 2.3 – *Histogramme des données de l'exemple 2.3.21. Le graphique de gauche donne les effectifs des classes de valeurs et celui de droite les fréquences.*

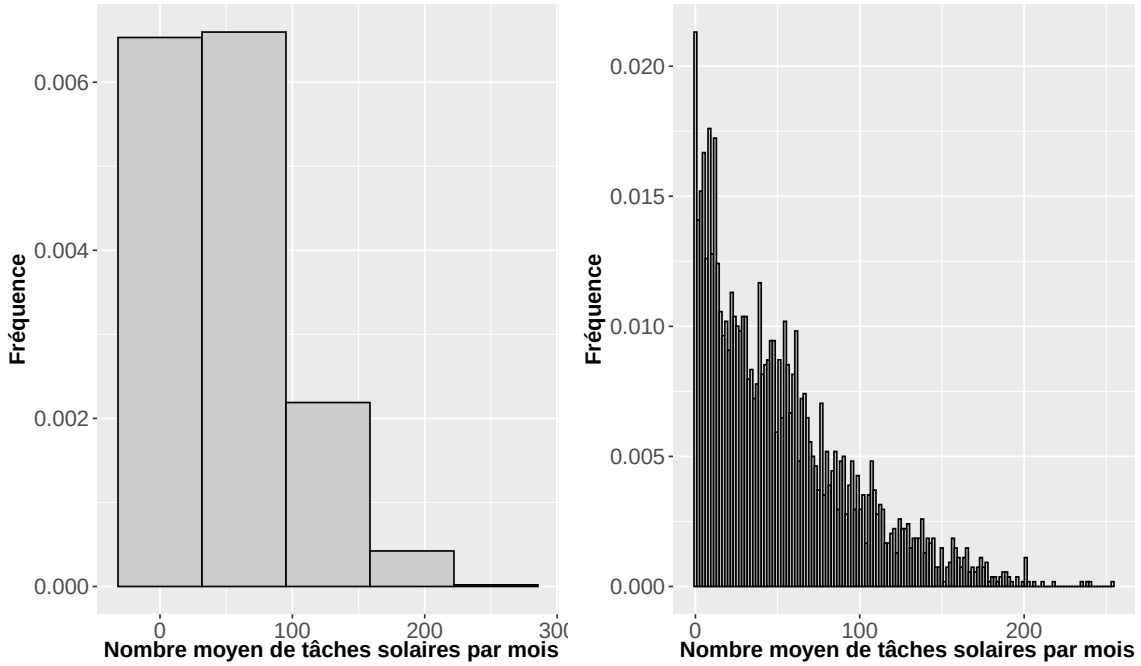


Figure 2.4 – *Histogramme peu parlants des données de l'exemple 2.3.21. Pour celui de gauche, il n'y a que trop peu de classes de valeurs et pour celui de droites il y en a trop.*

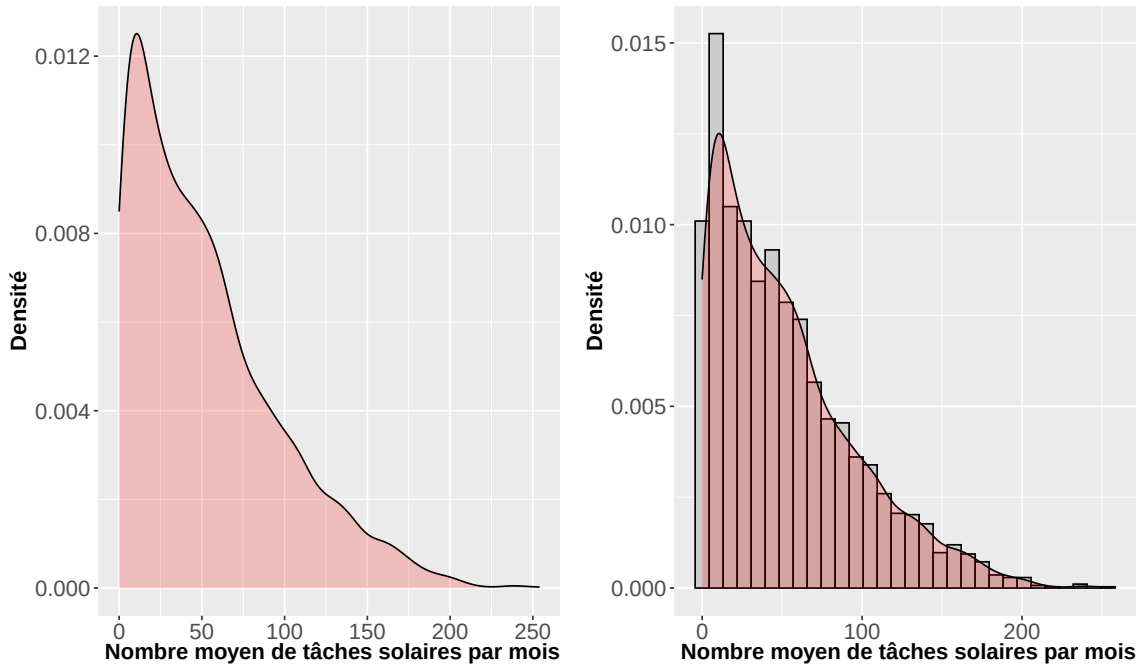


Figure 2.5 – *Densité de probabilité empirique pour les données de l'exemple 2.3.21. Le graphique de gauche donne la densité de probabilité empirique seule, et le graphique de droite la superpose avec l'histogramme des fréquences.*

Province	Fertilité	Agriculture
Courtelary	80.20	17.00
Delemont	83.10	45.10
Franches-Mnt	92.50	39.70
Moutier	85.80	36.50
Neuveville	76.90	43.50
Porrentruy	76.10	35.30
Broye	83.80	70.20
Glane	92.40	67.80
Gruyere	82.40	53.30
Sarine	82.90	45.20
Veveyse	87.10	64.50
Aigle	64.10	62.00
Aubonne	66.90	67.50
Avenches	68.90	60.70
Cossonay	61.70	69.30
Echallens	68.30	72.60
Grandson	71.70	34.00
Lausanne	55.70	19.40
La Vallee	54.30	15.20
Lavaux	65.10	73.00
Morges	65.50	59.80
Moudon	65.00	55.10
Nyone	56.60	50.90
Orbe	57.40	54.10

Province	Fertilité	Agriculture
Oron	72.50	71.20
Payerne	74.20	58.10
Pays d'en haut	72.00	63.50
Rolle	60.50	60.80
Vevey	58.30	26.80
Yverdon	65.40	49.50
Conthey	75.50	85.90
Entremont	69.30	84.90
Herens	77.30	89.70
Martigwy	70.50	78.20
Monthey	79.40	64.90
St Maurice	65.00	75.90
Sierre	92.20	84.60
Sion	79.30	63.10
Boudry	70.40	38.40
La Chauxdfnd	65.70	7.70
Le Locle	72.70	16.70
Neuchatel	64.40	17.60
Val de Ruz	77.60	37.60
Val de Travers	67.60	18.70
V. De Geneve	35.00	1.20
Rive Droite	44.70	46.60
Rive Gauche	42.80	27.70

Exercice 2.3.3 (Histogramme)**.** Faites un histogramme des fréquences pour chacune des deux variables des données de l'exercice 2.3.2, avec des classes de valeurs de longueurs 5.

Exercice 2.3.4 (Calcul incrémental)**.** Montrez l'équation suivante qui donne une relation incrémentale entre $\overline{x}_n$ et $\overline{x}_{n+1}$:

$$\overline{x}_{n+1} = \frac{n}{n+1}\overline{x}_n + \frac{1}{n+1}x_{n+1}.$$

2.4 Données atypiques

Dans le cadre d'une analyse descriptive, un aspect important à ne pas manquer consiste à déterminer s'il y a des données atyiques dans la base de données. Les données atypiques passent généralement sous le radar des approches statistiques standards, et des indicateurs les plus communs, parce qu'il est souvent plus intéressant et pertinent de faire ressortir une information concernant un effet "moyen" à propos d'un échantillon. En effet, lorsqu'on étudie un échantillon à des fins de généralisation à la population, on se focalise de manière légitime à décrire les points communs les plus partagés des individus, ainsi qu'à des profils standards observés dans l'échantillon. Pour une étude d'opinion, il est par exemple plus pertinent de faire ressortir que telle tranche d'âges ou telle classe socio-professionnel a une forte tendance à adhérer à telle valeur politique (ce qui correspond à l'élaboration un profil moyen) plutôt que de présenter les quelques cas de personnes qui sont en totale opposition avec les valeurs politiques partagées par leurs entourages respectifs (cas particulier).

Cependant, la présence de données atypiques dans une base de données peut avoir un effet non-négligeable (et potentiellement néfaste) sur l'analyse statistique. Ne pas les détecter et ne pas les traiter, revient à s'exposer à des problèmes de généralisation des résultats, voir la section 2.5.1 pour un exemple illustrant cela. De plus, dans le cadre d'une analyse statistique dans le domaine de la cybersécurité, les données atypiques sont souvent au cœur de l'intérêt. Les individus dont les mesures sont atypiques sont des candidats à étudier lorsqu'on cherche à détecter des fraudes ou des attaques sur un système informatique. Pour le formuler autrement, une donnée atypique c'est la bizarrerie qu'il faut être capable de détecter pour déceler un problème de sécurité.

A noter que détecter des données atypiques n'est pas une procédure simple puisqu'il n'existe aucune certitude absolue concernant le fait qu'une donnée soit effectivement atypique ou non, seulement certains principes peuvent nous guider dans cette approche. De plus, plusieurs méthodes de détection existent, mais il n'y a pas de raison de penser qu'une méthode est optimale, ou meilleure qu'une autre dans tout les cas de figure.

L'appellation de données atypiques n'est en soi pas évidente puisque suivant les sources, les définitions ne sont pas tout à fait les mêmes. En particulier, l'appellation de données atypiques peut englober plusieurs formes d'atypicité, et celles-ci ne sont parfois même pas considérées ou distinguées (suivant si le contexte le nécessite).

Définition 2.4.1 (Donnée atypique) *Une donnée atypique est une observation dont les caractéristiques différent de celles de la majorité des données.*

Parmi les données atypiques, on distingue dans ce document trois sous-types : les données extrêmes, les données anormales et les données aberrantes. Ces trois sous-types ne sont pas parfaitement distincts les uns des autres, mais suivant le contexte, on peut avoir un intérêt à s'intéresser à l'un d'entre eux.

Définition 2.4.2 (Donnée extrême) *Une donnée extrême est une observation dont la valeur sort largement du domaine des observations standards.*

Remarque 2.4.3 (Domaine des observations standards)**.** *Ce domaine correspond à un intervalle, à déterminer dans chaque cas de figure, avec des méthodes détaillées ci-dessous. Dès lors qu'une donnée n'est pas dans cet intervalle, on la désigne comme une donnée extrême.*

Exemple 2.4.4 (Crue d'une rivière)**.** *Le comportement standard d'une rivière est d'être autour d'un certain niveau (en terme de hauteur), mais il arrive qu'une conjonction d'événements provoque une crue. Le niveau de la rivière est alors extrêmement élevé. Le niveau de la rivière en crue correspond ici à une mesure extrême.*

Définition 2.4.5 (Donnée anormale) *Une donnée anormale ou anomalie est une observation qui ne correspond pas à une norme établit par un modèle (mathématique).*

Remarque 2.4.6 (Norme établie par un modèle)**.** *Cette norme correspond à un domaine, comme pour la remarque 2.4.3, mais dont la forme est potentiellement plus complexe qu'un intervalle, et pour lequel il est nécessaire d'employer des méthodes plus complexes pour le déterminer. Ces méthodes ne sont d'ailleurs pas au programme de cette ressource pédagogique, et seuls quelques exemples simples sont donnés dans ce document.*

Exemple 2.4.7 (Taille et volume d'un éléphant)**.** *Le poids d'un éléphant dépend en grande partie de sa taille (et donc de son âge). Il y a une fonction qui permet d'avoir une approximation du poids d'un éléphant en fonction de sa taille, bien qu'il y ait pour chaque éléphant une variation statistique autour du poids calculé. Par exemple, un éléphant mesurant 2 mètres (au garrot) peut peser 3 tonnes, et un autre mesurant 4 mètres peut peser 7 tonnes. Il n'y a donc rien d'atypique dans le fait de peser une éléphant qui fasse 3 ou 7 tonnes. Cependant, ce serait obtenir une donnée anormale que de mesurer un poids de 3 tonnes pour un éléphant de 4 mètres. Cet éléphant dans ce cas ne semble pas correspondre à la norme établie concernant la relation entre la taille et le poids d'un éléphant.*

Définition 2.4.8 (Donnée aberrante) *Une donnée aberrante est une observation qui correspond à une erreur de mesure ou d'échantillonnage.*

Exemple 2.4.9 (Température)**.** *Supposons qu'on collecte par soi-même la température à l'extérieur de son logement (fenêtre, balcon ou terrasse) à l'aide d'un thermomètre. Une des mesures qu'on obtient nous paraît bizarre puisqu'elle indique 75 degrés alors que la température pendant le reste de la journée n'a pas dépassé 30 degrés. Au moment de la mesure, on a noté que la fenêtre d'un des voisins renvoyée la lumière du soleil pile sur le thermomètre pendant quelques minutes. C'est la raison pour laquelle le thermomètre indique cette température, et cela correspond à une erreur de mesure, et donc à une donnée aberrante.*

Exemple 2.4.10 (Salaires dans une entreprise)**.** *Dans une entreprise d'une vingtaine de salariés, on relève les salaires de chacune des personnes pour étudier le salaire auquel pourrait prétendre un nouvel employé. Il s'avère que les salaires ne dépassent pas 2500 euros par mois, sauf pour le directeur qui culmine à environ 10000 euros par mois. La collection de données en question n'a pas de raison de contenir le directeur qui, d'un point de vue de la rémunération, n'est pas un employé comme les autres et qui n'aide pas à déduire quel devrait être le salaire d'un futur employé (qui ne sera pas un directeur). Il s'agit d'une erreur d'échantillonnage, on aurait pas du inclure le directeur dans l'échantillon, et cela correspond donc à une donnée aberrante.*

Bien qu'il soit important de savoir distinguer les différentes formes d'atypicité des données pour interpréter correctement le problème, le plus important est de savoir comment détecter une donnée atypique.

2.4.1 Détection pour une variable qualitative

Lorsqu'on dispose de données qualitatives, il suffit d'étudier la distribution empirique des modalités pour détecter des données atypiques. Dans ce cas, une donnée est atypique dès lors qu'il s'agit d'une unité statistique possédant une modalité qu'une très faible partie de l'échantillon possède aussi. Par exemple, l'hétérochromie (avoir les yeux vairons) est un trait physique rare (environ 1 personne sur 50000), et la mesure de la couleur des yeux d'une personne ayant une hétérochromie correspond à une mesure atypique. Pour un exemple contraire, il y a environ 10% de la population qui sont des gauchers. Cette fréquence est beaucoup trop élevée pour considérer que la modalité "être gaucher" puisse être considérée comme atypique.

Pour détecter une donnée atypique il faut se donner un seuil de fréquence en-dessous duquel on considérera un donnée comme atypique. Il n'y a pas de manière absolue de fixer ce seuil, et cela induit un caractère arbitraire dans la détection de données atypiques. Quoi qu'il en soit, dans certains contextes, on peut avoir une intuition concernant ce que devrait être ce seuil. Plus généralement il convient de 1) fixer ce seuil de tel sorte que très peu, ou pas, de données soient considérées comme atypiques et 2) ne pas se fixer un seuil plus grand que 1%, voire 5%.

2.4.2 Détection pour une variable quantitative

Pour ce qui est des données quantitatives, il a plus d'approches possibles pour déterminer des données atypiques. Voici ci-dessous une liste des méthodes les plus simples :

- **Seuil naturel :** Il y a des contextes pour lesquels on a une connaissance a priori suffisante pour nous-même déterminer ce qui peut être ou non atypique. Par exemple, pour ce qui est de la température (pour laquelle nous sommes experts puisque nous l'expérimentons tout les jours), il est possible de fixer un seuil de 45 ou 50 degrés au-delà duquel on considèrera qu'une mesure de température sera atypique.
- **Boxplot :** Comme vue dans la section 2.3.3.1, le graphique du boxplot permet d'estimer ce que pourrait être les bornes du domaine d'observation standard des mesures. On peut donc utiliser ce graphique pour déterminer les données atypiques. Les données relatives aux points qui "sortent" du boxplot peuvent être considérées comme des données atypiques. A noter que cette approche repose sur une hypothèse que les mesures suivent une loi normale (voir la ressource pédagogique de probabilité) et si les données suivent bien une loi normale alors le domaine estimé par l'intervalle du boxplot recouvre 99.3% des données possiblement échantillonnables. Autrement dit, si on obtient des points qui "sortent" du boxplot, soit il s'agit de données atypiques, soit l'hypothèse de normalité des données est incorrecte.
- **Seuil basé sur un quantile :** Comme pour l'approche décrite en section 2.4.1 pour les données qualitatives, on peut se fixer un seuil de fréquence s et il convient alors de calculer le quantile empirique au niveau $\hat{q}_s$ ou $\hat{q}_{1-s}$ (ou les deux) pour déterminer une (ou des) borne au-delà de laquelle on détermine les données comme atypiques.
- **Sensibilité d'un estimateur :** Cette approche un peu plus complexe consiste à évaluer à quel point l'estimateur d'intérêt dans le contexte est sensible à la suppression de chacune tour à tour des données. Cette méthode est partiellement employée dans la section 2.5.1. L'idée est que si une donnée a un impact

plus important que les autres données sur un estimateur, il s'agit probablement d'une donnée atypique. D'autres méthodes plus complexes sont au programme des prochaines ressources pédagogiques dédiées à la statistique, avec notamment l'utilisation de méthodes de classification, des modélisations des données et des méthodes de détection de rupture.

2.4.3 Traitement des données atypiques

Une fois qu'une donnée a été détectée comme atypique, voici les traitements possibles qu'on peut appliquer :

- **Suppression :** Une possibilité (qui n'est pas recommandée) est de suppression la donnée atypique. Cela peut se justifier, surtout s'il s'agit d'une donnée aberrante, mais il faut garder à l'esprit que dans de nombreux contextes, la collecte de données peut être coûteuse, difficile ou peut ne fournir qu'un nombre très limité de données. Il peut être alors difficile de se permettre de supprimer une donnée.
- **Imputation :** Méthode consistant à remplacer une valeur atypique par une autre valeur plus cohérente. Les possibilités de remplacement sont par la moyenne, la médiane, la moyenne des données les plus "proches" (au regard d'une autre variable), une borne définie par une raison dépendante du contexte, ou la valeur d'un quantile.
- **Etude séparée :** Mettre de côté les données atypiques et réaliser une étude séparée de sorte à ne pas laisser les données atypiques pertubrer les résultats de l'analyse statistique qui peut viser à déterminer des profils moyens, et d'analyser isolément les données atypiques afin de comprendre les caractéristiques communes de ces données.

Une autre possibilité consiste à adapter l'analyse statistique de sorte à utiliser des méthodes ou des indicateurs qui soient robustes à la présence de données atypiques. Cela a pour objectif de ne pas supprimer ou modifier une quelconque donnée.

Exemple 2.4.11. *On étudie la température moyenne pendant le mois de janvier de 1960 à Vancouver (Canada), dont les mesures sont ci-dessous :*

Jour	Température	Jour	Température	Jour	Température
jan01	1.70	jan11	3.80	jan21	4.40
jan02	2.30	jan12	3.30	jan22	4.00
jan03	1.10	jan13	2.80	jan23	3.60
jan04	3.60	jan14	1.80	jan24	1.30
jan05	1.90	jan15	5.20	jan25	3.90
jan06	0.60	jan16	10.70	jan26	2.90
jan07	2.30	jan17	3.50	jan27	3.20
jan08	3.30	jan18	4.40	jan28	2.10
jan09	3.60	jan19	4.10	jan29	2.70
jan10	2.40	jan20	4.30	jan30	3.30
				jan31	4.90

On s'interroge sur la potentielle atypicité de la donnée du 16 janvier. On calcule le quantile empirique à 99% *et on trouve* $\hat{q}_{99\%} = 9.026$. *Comme la données du 16 janvier dépasse ce seuil, on la catégorise comme étant une donnée atypique. Etant la valeur observée de la donnée atypique, on peut suspecter qu'il s'agit d'une donnée extrême : ce jour-là, il a fait très chaud. Dans ce contexte, où la dynamique des températures est assez régulière, on décide de réaliser une imputation de données en remplaçant la valeur atypique par la moyenne des températures observées les jours d'avant et d'après (trois jours avant et trois jours après). On obtient la valeur de* 4.76, *qui devient la nouvelle valeur du 16 janvier pour le reste de l'analyse.*

Mise en pratique des notions de la section 2.4

Exercice 2.4.1 (Identifier l'atypicité)**.** Pour les exemples ci-dessous, dites de quels types de données atypiques il s'agit (plusieurs types peuvent être utilisés) :

- Un tremblement de terre a été mesuré à 10.2 sur l'échelle de Richter.
- L'échelle de Scoville sert à mesurer l'intensité de force des piments, allant de 0 à 16 milliards, mais le piment le plus fort culmine à environ 3 milliards. Les mesures concernant le poivron oscillent entre 0 et 100. Après de nouvelles mesures, on s'interroge sur une mesure étrange de 2.3 milliards pour un poivron ordinaire.
- L'échelle de Schmidt permet de comparer la pénibilité des piqûres d'insectes, allant de 0 à 4. Pour une piqûre d'abeille, on trouve une étrange mesure de -2 pour laquelle on se pose la question de l'atypicité.
- Dans une étude concernant l'évolution de la mortalité infantile en France, on trouve une donnée étrange concernant un individu de 5 ans faisant 1.68 mètre.

- Une étude sensorielle est menée auprès d'un panel de consommateurs pour évaluer les profils de saveur de certains produits. Pour chaque produit, on demande aux consommateurs de mettre une note entre 0 et 10 concernant l'intensité de salé ressentie lors de la dégustation du produit. Deux données semblent étranges puisque pour l'une d'entre elles il s'agit d'un -6.7 et pour l'autre il s'agit d'un 0 alors que tout les autres consommateurs ont mis une note entre 7 et 9 pour ce produit.

Exercice 2.4.2 (Agriculture suisse)**.** Pour les données de l'exercice 2.3.2, catégorisez les données de la variable `Agriculture` en des classes de longueur 10%, puis déterminez s'il y a des modalités atypiques.

Exercice 2.4.3 (Fertilité Suisse)**.** Déterminez s'il y a des données atypiques dans la variable `Fertilité` des données de l'exercice 2.3.2.

2.5 Exemples de quelques biais statistiques

Dans cette section, quelques exemples sont donnés pour illustrer des problèmes fréquents qu'on peut rencontrer lorsqu'on analyse des données et qu'on cherche à en dégager une interprétation. Ces problèmes (qu'on appelle biais) reposent sur des visions contre-intuitives et sur des méprises sur ce que signifient et impliquent certaines notions en statistiques.

2.5.1 Moyenne et médiane

Voici ci-dessous les données de l'exemple 2.4.10, concernant les salaires des employés d'une entreprise :

	Statut	Salaire
1	Tech	1358.37
2	Tech	1317.76
3	Tech	1313.27
4	Ing	1879.94
5	Dir	9760.00
6	Tech	1401.77
7	Tech	1298.47
8	Tech	1340.13
9	Cad	2279.57
10	Ing	1854.05
11	Tech	1308.01
12	Tech	1288.63

	Statut	Salaire
13	Tech	1360.30
14	Ing	1916.50
15	Tech	1328.34
16	Cad	2221.00
17	Tech	1323.82
18	Tech	1339.26
19	Tech	1310.17
20	Tech	1360.44
21	Tech	1338.80
22	Ing	1969.60
23	Ing	1961.37

On note $x_1, x_2, \ldots, x_n$ les données des employés (non-directeur) et par x_{n+1} la donnée relative au directeur. On trouve les moyennes $\overline{x}_n = 1548.753$ et $\overline{x}_{n+1} = 1905.764$. La seule donnée du salaire du directeur a fait augmenter la moyenne d'un facteur de 1.230515 ($= \frac{\overline{x}_{n+1}}{\overline{x}_n}$). Cette sensibilité de la moyenne aux valeurs atypiques est justifiée mathématiquement par la proposition 2.5.1.

Proposition 2.5.1 (Sensibilité de la moyenne) *Pour un échantillon de n données, notées $x_1, x_2, \ldots, x_n$, on calcule la moyenne $\overline{x}_n$. Si on observe une nouvelle mesure x_{n+1}, pour qu'on ait la nouvelle moyenne $\overline{x}_{n+1}$ qui vérifie $\overline{x}_{n+1} > a\overline{x}_n$ pour un $a > 0$ donné, il faut que la nouvelle mesure vérifie $x_{n+1} > \overline{x}_n(a + n(a-1))$.*

Pour ce qui est des médianes, on obtient les résultats suivants : $m_n = 1349.25$ et $m_{n+1} = 1358.37$, et on constate effectivement que la prise en compte du salaire du directeur n'a que très peu modifié la médiane. De plus, pour illustrer cette robustesse aux valeurs atypiques, au contraire de la moyenne, la proposition 2.5.2 donne une majoration de la variation de la médiane par rapport à une nouvelle donnée.

Proposition 2.5.2 (Robustesse de la médiane) *Pour un échantillon de n données, notées $x_1, x_2, \ldots, x_n$, on calcule la médiane m_n. On note d la distance maximale entre deux données consécutives, à savoir $d = \max_{i=1,\ldots,n} x_{(i+1)} - x_{(i)}$. On a alors que quelque soit la valeur d'une nouvelle mesure obtenue, la nouvelle médiane m_{n+1} vérifie que $m_{n+1} \in [m_n - d, m_n + d]$.*

De plus, la proposition 2.5.3 donne les conditions d'observation de nouvelles données pour obtenir une augmentation de la médiane d'un facteur donné.

Proposition 2.5.3 (Faible sensibilité de la médiane) *Pour un échantillon de n données, notées $x_1, x_2, \ldots, x_n$, on calcule la médiane m_n. Pour un $a > 0$ donné, on note n_0 le nombre d'observations inférieures à $a \times m_n$ et n_1 le*

nombre d'observation supérieures à $a \times m_n$. Si on observe p nouvelles mesures $x_{n+1}, \ldots, x_{n+p}$, pour qu'on ait la nouvelle médiane m_{n+p} qui vérifie $m_{n+p} > am_n$, il faut que 1) $p > n_0 - n_1$ *et que* 2) *les nouvelles mesures soient toutes plus grandes que $a \times m_n$.*

Par exemple, pour obtenir une augmentation de la médiane d'un facteur de 1.230515, à savoir de passer de $m_n = 1349.25$ à $m_{n+p} = 1660.272$, il faudrait observer dans ce contexte $p = 8$ nouvelles données qui soient toutes supérieures à 1660.272.

2.5.2 Paradoxe de Simpson

Concernant l'évolution de la pandémie de Covid19 et l'efficacité des vaccins, Israël est un pays qui est surveillé de part sa stratégie vaccinale particulièrement intense et rapide dès la mise à disposition des vaccins. A la mi-août 2021, de nouvelles données rendues disponibles par le gouvernement israélien ont laissé penser que le vaccin n'était pas efficace contre les formes sévères de la maladie. Les chiffres en question sont les suivants :

Nombre de cas sévères actuellement hospitalisés	
Personnes non vaccinées	Personnes vaccinées
214	301

La présentation de ces résumés statistiques expose le lecteur a une conclusion erronée, de part un mauvais choix dans la manière de réduire l'information, qui fait apparaître le paradoxe de Simpson. Ce paradoxe implique qu'il n'est pas possible, au vu de ces résumés, de statuer quant à un effet potentiel ou non du vaccin sur l'apparition de formes sévères de la maladie.

Pour comprendre ce paradoxe, il faut rendre disponible des données supplémentaires : les tailles des sous-populations des personnes vaccinées et des personnes non-vaccinées.

	Statut vaccinal		Nombre de cas sévères actuellement hospitalisés	
	Personnes non vaccinées	Personnes vaccinées	Personnes non vaccinées	Personnes vaccinées
Effectif	1302912	5634634	214	301
Fréquence	18.78%	81.22%	16.42	5.34

Pour les fréquences du nombres de cas sévères, il s'agit de la fréquence parmi les personnes ayant le même statut vaccinal pour 100k individus (unité standard pour mesurer la présence d'une maladie contagieuse dans une population). On constate ici que les personnes étant vaccinées ont moins de propension à contracter une forme sévère de la maladie. Le paradoxe de Simpson ici à l'œuvre fait oublier que même si une partie de la population (les personnes vaccinées) ont moins de chance de faire des formes sévères, dès lors que cette partie de la population est beaucoup plus importante que l'autre, cette partie de la population peut apporter plus de personnes malades que le reste de la population non-vaccinées.

Pour aller plus loin, voici un résumé encore plus complet de la situation, qui permet de constater l'efficacité du vaccin en fonction de la classe d'âge :

	Pourcentage des personnes		Cas sévères pour 100k individus	
âge	Non-vaccinées	Vaccinées	Non-vacciné	Vacciné
12-15	70.10	29.90	0.30	0.00
16-19	26.50	73.50	1.60	0.00
20-29	23.80	76.20	1.50	0.00
30-39	19.10	80.90	6.20	0.20
40-49	15.60	84.40	16.50	1.00
50-59	12.00	88.00	40.20	2.90
60-69	10.20	89.80	76.60	8.70
70-79	5.40	94.60	190.10	19.80
80-89	7.40	92.60	252.30	47.90
90+	9.50	90.50	510.90	38.60

On peut de nouveau constater que le nombre de cas sévères chez les personnes vaccinées est inférieur à celui des personnes non-vaccinées, et ce quelque soit l'âge. Pour conclure, certes il y a plus de malades chez les personnes vaccinées, mais c'est parce qu'il y a beaucoup plus de personnes vaccinées. Cependant, en proportion il y a plus de personnes malades non-vacinnées, ce qui semble plutôt aller dans le sens de l'efficacité du vaccin.

2.5.3 Groupes séparés et incohérence des indicateurs de position

Bien que cela puisse paraître contre-intuitif, il y a des cas pour lesquelles un indicateur tel que la moyenne ne représente aucun individu. Prenons par exemple le cas des mesures de tâches solaires, voire l'exemple 2.3.21 mais en prenant les données relatives aux années 1800 et 2000. Les données en question sont les suivantes :

Moyenne mensuelle	Année	Mois
6.90	1800	Jan
9.30	1800	Feb
13.90	1800	Mar
0.00	1800	Apr
5.00	1800	May
23.70	1800	Jun
21.00	1800	Jul
19.50	1800	Aug
11.50	1800	Sep
12.30	1800	Oct
10.50	1800	Nov
40.10	1800	Dec

Moyenne mensuelle	Année	Mois
90.10	2000	Jan
112.90	2000	Feb
138.50	2000	Mar
125.50	2000	Apr
121.60	2000	May
124.90	2000	Jun
170.10	2000	Jul
130.50	2000	Aug
109.70	2000	Sep
99.40	2000	Oct
106.80	2000	Nov
104.40	2000	Dec

La moyenne des ces données mensuelles est 67.00417, or comme on peut le constater dans les tableaux de données précédents, autour de la valeur de la moyenne il n'y a aucune mesure observée. Cela peut paraître contre-intuitif puisqu'on a l'habitude des répartitions de données avec "un seul tas de données" (on parle plutôt d'un seul groupe ou de distribution unimodale) qui se répartissent autour de la moyenne. Cependant, s'il y a deux groupes de données qui se distinguent bien l'un de l'autre (on parle de distribution bimodale), la moyenne peut se trouver exactement entre les deux groupes, proche d'aucune donnée. Dans ce cas-là, l'utilisation de la moyenne, telle quelle, n'est pas un bon moyen d'avoir un indicateur parlant concernant la distribution des données. Renseigner la moyenne dans ce type de cas est une erreur d'analyse descriptive.

Si les données se répartissent en deux groupes ou plus, pour réaliser une analyse descriptive cohérente, il faut calculer les résumés statistiques pour chacun des groupes de données. Pour cet exemple, on obtient alors les moyennes suivantes : 14.475 pour l'année 1800, et 119.5333 pour l'année 2000.

2.6 Diapos de cours et exercices de travaux dirigés

Les pages suivantes (jusqu'au début du chapitre 3) sont les versions "prises de notes" des diapos de cours et les feuilles de travaux dirigés qui sont au programme de cette ressource pédagogique.

Les chapitre 1 des diapos couvrent l'introduction (voir le chapitre 1) et l'organisation de cette ressource pédagogique. Les chapitre 2 des diapos concernent les notions d'échantillonnnage et des notions générales à retrouver dans la section 2.1. Pour le chapitre 3 des diapos, il faut trouver le parallèle avec la section 2.2 et la section 2.3 pour le chapitre 4 des diapos. Le dernier chapitre des diapos concerne la section 2.4.

La feuille de TD 1 est relative à la section 2.1, la feuille de TD 2 aux sections 2.2 et 2.3, et la feuille de TD 3 à la section 2.4.

Chap. 1 – **Introduction : Descriptive ?**

Statistique et objectif

L'objectif des approches statistiques est de

Elles ont un intêret lorsque

La procédure consiste à

L'exemple du procès

La situation se transpose à l'exemple d'un procès qui se tient pour

La complexité de la situation tient à l'incertitude concernant

Le procès se déroule de la manière suivante

Le réel, l'observation et l'aléatoire

Les approches statistiques servent généralement à

Le réel est accessible uniquement par le bias de

Postulat :

Les approches statistiques sont entre autres basées sur l'aléatoire, ce qui permet de modéliser

La démarche

Les principales étapes de la procédure d'une étude statistiques sont :

Concevoir

Collecter et stocker

Pré-analyser

Inférer

Valider

Décider

Statistique descriptive : 1$^{\text{ère}}$ moitié

La dénomination de "statistique descriptive" comprend :

A la fin de ce module vous devrez être capable de :

Statistique descriptive : 2$^{\text{nd}}$ moitié

La dénomination de "statistique descriptive" comprend :

A la fin de ce module vous devrez être capable de :

La ressource RES 1-05

Objectifs de cette ressource :

1$^{\text{ère}}$ moitié de la ressource
(Paul-Marie Grollemund)

- Comprendre ce que sont des données
- D'où elles viennent
- Ce sur quoi elles peuvent nous éclairer
- Représentations graphiques et numériques
- Détecter des données atypiques

2^{nd} moitié de la ressource
(Maeva Paradis)

- Comprendre ce que sont des données
- D'où elles viennent
- Ce sur quoi elles peuvent nous éclairer
- Représentations graphiques et numériques
- Détecter des données atypiques

Enseignants à contacter :
en cas de problèmes ou de questions

Se rendre au bureau des enseignants STID ou par mail à :
paul_marie.grollemund@uca.fr

maeva.paradis@uca.fr

Déroulement des enseignements

Cours Magistraux (CM)
5 séances de CM
Semaines 37, 38, 39, 40 et **45** (**examen**)
Enseignant : Paul-Marie Grollemund

Travaux dirigés (TD)
7 séances de TD
Semaines 38, 39, 40, 41, **42** (**examen**) et 43
Enseignant groupe 1 : Paul-Marie Grollemund
Enseignant groupe 2 : Paul-Marie Grollemund

Travaux pratiques (TP)
3 séances de TP
Semaines 39, 40 et **43** (**exeman**)
Enseignant groupe A : Paul-Marie Grollemund
Enseignant groupe B : Paul-Marie Grollemund
Enseignant groupe C : Paul-Marie Grollemund
Enseignant groupe D : Ousmane Cissé

Examens
A chaque début de CM, un contrôle continu (à compter du prochain CM)
Examen de CM (individuel) pendant la semaine 45
Examen de TD (en groupe) pendant la semaine 42
Examen de TP (individuel) pendant la semaine 43
Note finale : moyenne des 3 meilleures des 4 notes

Chap. 2 – **Échantillonnage**

Exemple introductif

Contexte :
Supposons qu'à l'IUT, il y ait 314 étudiants inscrits.

Les enseignants se demandent combien de temps les étudiants passent à préparer les examens. Une solution serait de demander aux étudiant :
"Combien de temps par semaine consacrez vous à vos loisirs ?"

Objectif :
Un des objectifs de ce questionnement est de faire le lien entre :
— "le temps consacré aux loisirs" et "le temps consacré à étudier",
— "le temps consacré à étudier" et "la note aux examens".

Problème :
Demander l'avis de tout les étudiants, et analyser l'ensemble des données, serait beaucoup trop long.

Comment répondre à cette question en y consacrant un temps raisonnable ?

Une solution de statisticien

Solution : D'un point de vue d'un statisticien il faudrait :
.
.
.
.

Problème :
Cette solution n'est pas miraculeuse et elle a aussi ces problèmes :
.
.
.
.
Mais on va d'abord avoir besoin d'introduire des notions importantes.

Notions importantes

Définition : Population
.
.
.

Définition : Echantillon
.
.
.

Définition : Unité statistique ou individu
.
.

Notions importantes (2)

Définition : Paramètre .
.

Définition : Variable ou caractère
.
.

Notions importantes (3)

Définition : Donnée ou observation
.
.

Définition : Quantité (variable) aléatoire
.
.

Définition : Statistique
.
.

Estimateur

Définition : Estimateur
.
.

Exemple : estimateur
L'enseignant interroge 91 étudiants.
Le **temps moyen** (consacré aux loisirs) calculé sur un échantillon de 91 étudiants est un estimateur du temps consacré aux loisirs pour l'ensemble de la population de 314 étudiants.

Exemple : application
Après interrogatoires et calculs, si l'enseignant détermine que les 91 étudiants interrogés attribuent **en moyenne 20h** à leurs loisirs par semaine, on peut s'attendre à ce que le temps moyen des étudiants de toute la promotion soit **proche de 20h** par semaine.

Problème :
Serait-ce très différent si on recommençait ?

Pour tenter de répondre à cette problématique, il est nécessaire d'introduire des notions permettant de comprendre et d'étudier la variabilité des résultats.

Quantité aléatoire ?

Définition : Echantillonnage aléatoire simple :
.
.
.

Remarque : Donnée et aléatoire
La donnée relative à un individu n'est pas aléatoire en elle-même.
Le fait d'avoir mesuré cette donnée, suite à un échantillonnage aléatoire, rend cette donnée aléatoire.

Tout ce qui dépend de l'échantillonnage est dès lors aléatoire.
Tout ce qui dépend des données est dès lors aléatoire.

Remarque :
Le choix des données de l'échantillon parmi la population, est un choix aléatoire.

Combien de scénarios d'échantillonnage ?

Exemple : (étudiants et loisirs)
Nbre de possibilités pour le premier individu choisi : 314
Nbre de possibilités pour le deuxième individu choisi : 313
Nbre de possibilités pour le troisième individu choisi : 312
⋮ ⋮ ⋮ ⋮ ⋮ ⋮
Nbre de possibilités pour le $91^{\text{ème}}$ individu choisi : 224

Nbre de scénarios possibles :
.
.
.
.
.
.
.

Notations : variable et données

Notation : X (action de mesurer)
On écrit X la variable aléatoire : "mesurer un individu quelconque".

Notation : X_i (action de mesurer)
On écrit X_1 la variable aléatoire : "mesurer un individu numéroté 1".
De la même manière, on a les notations $X_2, X_3, \ldots$ et on utilisera souvent X_i pour la variable aléatoire relative au fait de mesurer l'individu i.

Notation : x_i (résultat de la mesure)
On écrit x_1 la mesure obtenue pour l'individu numéroté 1.
De la même manière, on a les notations $x_2, x_3, \ldots$ et on utilisera souvent x_i pour la mesure de l'individu i.

Notation : y_i
On écrit y_1 la première mesure possible dans la population.

De la même manière, on a les notations $y_2, y_3, \dots$ et on utilisera souvent y_i pour la $i^{\text{ème}}$ mesure possible dans la population.

Remarque :
Dans la suite, on fera parfois la confusion entre les notations X_i et x_i.

De plus, on n'utilisera que rarement la notation y_i. Cette notation ne sera utilse pour discuter des valeurs possibles dans la population, et donc d'étudier les propriétés des résultats avec échantillonnage.

Espérance

Définition : Espérance
.
.

Remarque :
À ne pas confondre avec la moyenne de l'échantillon.
(Il existe un lien étroit entre ces deux notions.)
La moyenne des données de l'échantillon est dite **empirique**
(parce qu'on la connait en pratique).
La moyenne de la population est dite **théorique**
(parce qu'on ne la connait jamais en pratique).

Interprétation :
.
.

Remarque :
Origine : L'espérance est le gain qu'on peut se permettre d'attendre si on joue successivement un grand nombre de fois à un même jeu (espérance de gain).

Propriété : pour calculer l'espérance
.
.
.

Espérance (2)

Propriété :
Certaines valeurs apparaissent plusieurs fois dans la population.
Par exemple, sur une population de 20 individus, si
- la valeur y_1 apparait 2 fois,
- la valeur y_2 apparait 6 fois,
- la valeur y_3 apparait 12 fois et
- qu'il n'y a pas d'autres valeurs,

alors l'espérance peut se calculer (par factorisation) de la manière suivante :
.
.
.

Propriété :
Autrement formulé, l'espérance est
.
.
.
où y_i est la $i^{\text{ème}}$ valeur possible parmi les p différentes valeurs possibles dans la population.

Remarque :
Ici $P(X = y_1)$ est la probabilité que la mesure d'un individu, pris au hasard, soit égale à la valeur y_1.

Opérateur d'espérance

Définition : Opérateur d'espérance
.
.
.
.

Propriété : (Formule)
.
.
.

Remarque :
En pratique, on utilisera parfois la terminologie "espérance" pour désigner l'opérateur d'espérance $\mathbb{E}$.
C'est un abus de langage qui introduit une confusion avec l'espérance μ mais qui est courrament utilisé.

Propriété :
L'opérateur d'espérance est *linéaire*, c'est-à-dire :
.
.
.
.

Espérance de la moyenne

Définition : Moyenne
.
.
.
.
.

Propriété :
Par linéarité de l'espérance, m'espérance de la moyenne est la moyenne des espérances.
.
.
.
.
.

Interprétation :
.
.
.

Importance des cas extrêmes ?

⋆ Cas particulier 1 : (chance)
La moyenne des données des 91 étudiants peut être en pratique exactement égale à μ.
Mais on ne peut pas le savoir en pratique.

⋆ Cas particulier 2 : (pas de chance)
La moyenne des données n'est pas du tout proche de μ.
Par exemple : la proportion d'étudiantes dans la promo :
.
.
.
.

⋆ Est-ce possible d'évaluer la probabilité que ces cas particuliers arrivent alors qu'il y a énormément de configurations possibles ?
Pour y répondre, on va avoir besoin d'introduire une notion de variance.

Variance

Définition : Variance
.
.
.

On note σ^2 la valeur de la variance de la population :

$$\sigma^2 = \frac{1}{N}\left[(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_N - \mu)^2\right]$$

Interprétation :
.
.
.

Opérateur de variance

Définition : : Opérateur de variance
.
.

Propriété : (Formule)
En utilisant la même réflexion que pour l'espérance, la factorisation des valeurs possibles de la population y_i permet d'écrire :

$$\mathbb{V}(X) = (y_1 - \mathbb{E}(X))^2\, P(X = y_1)^2 + \cdots + (y_N - \mathbb{E}(X))^2\, P(X = y_N)$$

où y_i est la $i^{\text{ème}}$ valeur possible dans la population.

Remarque :
En pratique, on utilisera parfois la terminologie "variance" pour désigner l'opérateur de variance $\mathbb{V}$.
C'est un abus de langage qui introduit une confusion avec la variance σ^2 mais qui est courrament utilisé.

Propriété :
Les propriétés de calculs de l'opérateur de variance sont :

$$\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + 2\,\mathrm{Cov}(X_1, X_2) + \mathbb{V}(X_2)$$
$$\text{Si } X_1 \perp\!\!\!\perp X_2 : \mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$$
$$\mathbb{V}(a \times X) = a^2 \times \mathbb{V}(X)$$

où a est une constante (un nombre fixé) et $\perp\!\!\!\perp$ indique que deux variables sont indépendantes l'une de l'autre.

Variance de la moyenne

Propriété : La variance de la moyenne
Pour une population de taille N et pour un échantillon de taille n, le calcul de la variance de la variable $\bar{X}$ donne :
.
.
.

Interprétation : Petit échantillon
.
.
.

Interprétation : Grand échantillon (presque aussi grand que la population)
.
.
.

Échantillonnage stratifié

Remarque :
Lorsqu'on réalise un échantillonnage aléatoire simple, on peut se retrouver avec une configuration extrême.

Exemple :
Lors d'une enquête d'opinion, on effectue un sondage sur un échantillon de 1000 personnes afin de déterminer le positionnement des gens par rapport à la cyber-criminalité. Imaginons qu'on se retrouve par "malchance" à n'avoir tiré aléatoirement que des personnes ayant entre 70 et 90 pour la constitution de l'échantillon.
On peut s'attendre à ce que cette configuration ne permettent pas de donner une image *représentative* de l'opinion de la population.

Solution :
Pour éviter les erreurs dues à l'échantillonnage, une autre méthode est possible :
l'échantillonnage stratifié

Définition : Strate
.
.
Les strates doivent être définies telles qu'elle soient :
.
.
.
.

Exemple : La région d'origine des étudiants.

Échantillonnage stratifié (2)

Définition : Echantillonnage stratifié
.
.

Propriété :
.
.

Remarque :
La stratification doit se faire en rapport à une quantité étudiée.

Exemple :
- Quantité d'intérêt : les intentions de vote
- Stratification non pertinente : par nombre de frère(s) et soeur(s)
- Stratification pertinente : par catégorie socio-professionnelle

Propriété : Généralisation
.
.
.
.

Un résumé

Un résumé rapide de ce chapitre :
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

TD 1 – Echantillonnage et variables

Exercice 1. Le but de cet exercice est que vous fassiez connaissance avec vos camarades. Vous serez amenés à vous lever, à interroger vos camarades et nous ferons un état des lieux des réponses que vous avez obtenu avec quelques outils statistiques.

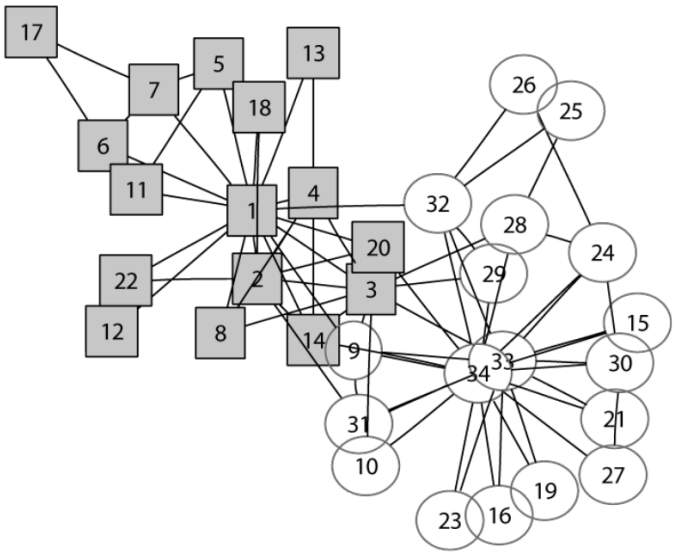
1. Vous avez 5 minutes pour vous lever et demander à 4 autres de vos camarades les informations suivantes : leurs numéros d'étudiant, leurs genres (femme ou homme), nombres de frère(s) et sœur(s), régions d'origine et leurs tailles (en cm).
2. Rassemblez par écrit leurs réponses sous une forme adéquate.
3. Identifiez quelle est la population, l'échantillon et l'unité statistique.
4. Déterminez les types des variables dont vous disposez.
5. Calculez les effectifs et les fréquences pour chacune des variables.
6. Que pensez-vous de la notion d'"effectif" pour les différentes variables ?
7. Calculez la moyenne pour chacune des variables. Ayez un œil critique sur ces calculs.
8. Rassemblons les données aux tableaux et calculons la variance de vos moyennes.
9. Déterminez si votre échantillon est représentatif (par rapport à chacune des variables).

Exercice 2. Une entreprise a relevé le nombre de requêtes qui étaient effectuées sur leur serveurs par crainte d'avoir une intrusion malveillante. Ci-dessous voici, sur plusieurs jours, le cumul de ces nombres de requêtes toutes les deux heures.

heures	0.00	2.00	4.00	6.00	8.00	10.00	12.00	14.00	16.00	18.00	20.00	22.00
jour 1	0	1	0	5	10	11	20	22	31	30	2	3
jour 2	3	4	0	6	9	8	19	17	30	30	0	1
jour 3	1	0	1	9	11	12	16	14	33	21	1	0
jour 4	0	9	10	4	8	13	20	21	35	31	2	1
jour 5	1	4	6	9	20	21	20	17	27	27	3	5
jour 6	9	1	1	2	0	1	0	1	2	0	1	0
jour 7	1	2	0	0	1	0	0	2	1	0	0	0

1. Déterminez quelle est la population, l'échantillon et l'unité statistique.
2. Quel est le type de la variable ?
3. Calculez les effectifs par jours.
4. Calculez les effectifs pour les plages horaires suivantes : 0.01-6.00, 6.01-12.00, 12.01-18.00 et 18.01-0.00.
5. Déterminez le jour moyen, les moyennes par plages horaires.
6. Expliquez les différences de moyennes que vous observez par plages horaires.
7. Déduisez-en un regroupement des jours. Recalculez les moyennes.
8. Déterminez la variance par plages horaires.
9. Expliquez les différences de variances.

Exercice 3. Entre 1970 et 1972, Wayne Zachary a étudié, en anthropologie sociale, un réseau social de petite taille avec l'objectif d'étudier la scission d'un groupe. Le groupe en question était un club de karaté qui s'est séparé en deux. Voici un graphe qui représente ce club.



Chacun des noeuds de ce graphe représente une personne. Les arrêtes représentent des liens d'amitiés et la couleur des noeuds indique dans quel sous-groupe la personne a fini après la scission. Dans la suite, le groupe A sera le groupe gris et le groupe B le blanc.

1. Quelle est l'unité statistique, la population et l'échantillon ?
2. Calculez les degrés des noeuds. Répartissez-vous les tâches.
3. Déterminez la moyenne et la variance des degrés.
   (Le degré d'un noeud d'un graphe est le nombre d'arrêtes reliées à ce noeud.)
4. Déterminez la moyenne des liens d'amitié entre les personnes du groupe A et celles du groupes B ($A \to B$).
5. Même question pour les liens $A \to A$ et $B \to B$. Répartissez-vous les tâches.
6. Calculez la proportion de liens d'amitié qui vont d'un groupe vers l'autre (parmi l'ensemble des liens existants).
7. Déterminez un sous-groupe représentatif du club de karaté et déterminez la proportion de liens d'amitiés qui vont d'un groupe vers l'autre. Dans ce contexte, un sous-groupe représentatif sera construit en prenant des individus des deux groupes en respectant les proportions des groupes.
8. Comparons vos résultats à la proportion totale calculée précédemment.

Exercice 4. Considérons un échantillon de 20 personnes pour lesquelles on dispose d'une « indication » sur leur intérêt respectif pour la politique (A = Aucun, B = Faible, C = Moyen, D = Important) :

Bernadette (C), Laurent (C), Annie (A), Abdellah (B), Martin (A), Mireille (B), Pierre (C), Maurice (B), Farida (B), Jacques (C), Leila (D), Carole (B), Didier (A), Monique (C), Noémie (C), Hélène (C), Gérald (C), Gérard (D), André (C), Wilfried (C).

1. Donnez la population, l'unité statistique, le caractère étudié (nature, modalités etc).
2. Quelle est la taille de la population ?
3. Construisez un tableau pour représenter la distribution statistique étudiée.
4. Représentez cette distribution à l'aide d'un graphique.

Chap. 3 – **Grandeurs et graphiques pour décrire une variable**

Type de données

Remarque :
Il y a plusieurs types de données possibles.
Suivant le type, on utilisera pas les même outils pour travailler les données.

Définition : Les différents types
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Le traitement est différent entre les données discrètes (1.-4.) et les continues (5.).

Dans ce chapitre, on va traiter le cas de données qualitatives.

Modalité et distribution

Définition : Modalité, classe ou niveau
.
.

Notation : Modalités
On note K le nombre de modalités possibles d'une variable qualitatives et $c_1, c_2, \ldots, c_K$ ces modalités.

Exemple : Lettre d'un mot de passe
Pour un mot de passe, identifier la première lettre est une variable qualitative, dont les modalités sont les lettres de l'alphabet. Il y a $K = 26$ modalités qui sont $c_1 = \texttt{a}, c_2 = \texttt{b}, \ldots,$ et $c_{26} = \texttt{z}$.

Remarque :
On peut supposer que pour un mot de passe donné (qu'on ne connait pas), chacune des lettres (modalités) a une probabilité théorique d'être utilisée pour la première lettre. On note p_k la probabilité de la $k^{\text{ème}}$ modalité.

Définition : Distribution d'une variable qualitative
.
.

Objectif :
Connaître la distribution d'une variable qualitative permet d'appréhender mathématiquement le caractère aléatoire de la variable qualitative.

Solution :
Observer des données et en déduire autant qu'on peut cette distribution.

Données et résumés

Notation : Données
Des données qualitatives concernant n individus.
On note les valeurs des données : $x_1, x_2, \ldots, x_n$.

Chaque donnée est une observation de la variable qualitative, et donc chaque donnée correspond à un modalité. Si la première observation du phénomène donne la $k^{\text{ème}}$ modalité, alors : $x_1 = c_k$.

Exemple :
Le premier mot de passe est `"squelette"` donc la première lettre est $x_1 = \texttt{s}$.

Définition : Résumé statistique
.
.

Exemple :
Indiquer que 62% des premières lettres sont des voyelles correspond à donner une résumé statistique.

Propriété :
.
.
.
.
.
.

Les résumés numériques et graphiques

Dans ce qui suit, on va introduire les résumés suivants :
Les résumés numériques : .
.
.
.
.

Remarque :
Ces outils servent à renseigner des aspects pertinents des observations concernant une variable qualitative.

Contexte :
Base de données de mot de passe : $\sim$ 340k

Problème :
Trop d'informations pour un humain
Besoin de résumés statistiques

Résumés numériques

Contexte :
Supposons disposer des données suivantes :

i	1	2	3	4	5	6	7	8	9	10
x_i	b	z	o	e	b	e	a	i	a	e

Définition : Effectif d'une modalité
.
.

Définition : Fréquence d'une modalité
.
.

Exemple :
On note $n_{\texttt{a}}$ l'effectif de la modalité `a` : $n_{\texttt{a}} = 2$.
On note $f_{\texttt{e}}$ la fréquence de la modalité `e` : $f_{\texttt{e}} = 3/10$.

Définition : Mode
.
.

Exemple :
On note m le mode est ici : $m = \texttt{e}$.

Distribution empirique

Définition : Distribution empirique
.
.

Exemple :

x_i	a	b	c	d	e	f	g	h	i	j	k	l	m
f_i	0.2	0.2	0	0	0.3	0	0	0	0.1	0	0	0	0
x_i	n	o	p	q	r	s	t	u	v	w	x	y	z
f_i	0	0.1	0	0	0	0	0	0	0	0	0	0	0.1

Remarque : Empirique et théorique
.
.
.
.
.
.
.
.
.

Indicateur graphiques (1)

.
.
.
.
.

En résumé

Un résumé rapide de ce chapitre
- Type de données
  - Distinguer parmi les qualitatives
- Notions autour des variables qualitatives
  - Modalité, distribution
- Résumé statistique
  - Qu'est-ce que résumé, réduction de l'information
- Les résumés numériques
  - Effectif, fréquence, distribution empirique
- Empirique et théorique
  - Lien et différences
- Les résumés graphiques
  - Les possibilités et comment choisir
- Savoir choisir des résumés

Chap. 4 – **Résumer une variable quantitative : Empreintes digitales**

Les résumés numériques et graphiques

Notation :
Des données quantitatives continues concernant n individus.
On note les valeurs des données : $x_1, x_2, \ldots, x_n$.

Définition : résumé statistique
.
.

Propriété :
Utiliser un résumé statistique revient à "réduire l'information" présente dans les données.
Réduire l'information peut être acceptable au profit d'obtenir une information interprétable.

Dans ce qui suit, on va introduire les résumés suivants :
Les résumés numériques : .
.
.
.
Les résumés graphiques : .
.
.
.

Exemple de données

Contexte :
Base de données d'empreintes digitales : 8 doigts de 10 individus

Objectif :
Un objectif possible, mais complexe, serait de vouloir générer une empreinte artificielle pour attaquer un système d'authentification biométrique.

Problème :
Etudier l'intensité des pixels pour avoir une première intuition de comment générer une empreinte.

Donner un aperçu synthétique de la base de données.

Résumés graphiques (1)

Définition : Catégorisation .
.
Remarque : Effectif .
.

Définition : Histogramme des effectifs
Correspond à une diagramme en barre pour les effectifs des classes de valeurs.
.
.
.
.

Résumés graphiques (1)

Définition : Histogramme des fréquences
.
.

Remarque :
Pour le graphique de la diapo précédente, la hauteur H_k de la $k^{\text{ème}}$ barre correspond à l'effectif de la $k^{\text{ème}}$ classe de valeur.

Pour le graphique ci-dessous, la hauteur H_k de la $k^{\text{ème}}$ barre est déterminée de sorte à ce que l'aire de chaque barre soit égale à la fréquence f_k.
Si la largeur de la $k^{\text{ème}}$ barre et ℓ_k, alors la hauteur est : $H_k = f_k/\ell_k$.

Propriété :
.
.
.
.
.
.

Résumés graphiques (2)

Définition : Densité de probabilité
.
.
.
.
.

Définition : Densité de probabilité empirique
.
.

Remarque :
En pratique, on ne peut pas facilement calculer la densité empirique et on utilise des fonctions faciles d'utilisation à partir d'un ordinateur (voir pendant les séances de TP), sans avoir besoin de comprendre des notions complexes.

Indicateurs de tendance centrale (1)

Notation :
On note la moyenne $\overline{x}$, ou $\overline{x}_n$ lorsqu'il est utile d'indiquer quelle est calculé à partir de n données.
Définition : Moyenne
.
.
.
.

Remarque : Moyenne et espérance
La moyenne est une version empirique de l'espérance. Autrement dit, on peut calculer une approximation de l'eséprance en calculant la moyenne.
Exemple :
Sur les données d'empreine digitales (80), on trouve en pratique la moyenne suivante pour un pixel donné : $\overline{x}_n = \frac{1}{80}\left(0.351 + 0.064 + \cdots + 0.408\right) \approx 0.161$

Propriété :
.
.

Indicateurs de tendance centrale (2)

Notation :
On note M la médiane d'un échantillon de données.
Définition : Série ordonnée
.
.
.

Définition : Mediane
.
.

Méthode de calcul :
.
.
.
.
.

Propriété :
Cette manière d'évaluer la tendance centrale est robuste à la présence de données extrêmement grandes.

Indicateurs de tendance centrale (3)

Définition : Mode
.
.

Propriété :
Les notions de mode pour les variables qualitatives ou quantitatives sont proches mais ils ne se calculent pas de la même manière.
Remarque :
Le calcul en pratique n'est pas évident puisqu'il faut calculer la densité de probabilité empirique, ce qui se fait par ordinateur avec des fonctions déjà existantes.

Indicateurs de dispersion (1)

Définition : Variance empirique
.
.
.
Théorème : Köning-Huygens
.
.
.
Remarque :
La variance empirique correspond à la moyenne des écarts (au carré) à la moyenne.
Définition : Variance empirique corrigée
.
.
.
Définition : Ecart-type (corrigé)
Indique la dispersion des données dans la même unité que les données : $s = \sqrt{s^2}$ et $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.

Indicateurs de dispersion (2)

Définition : Coefficient de variation
.
.
Définition : Coefficient de variation empirique
.
.
Remarque :
A n'utiliser qsue si les données sont que positives ou que négatives, et si la moyenne n'est pas proche de 0.

Indicateurs de dispersion (3)

Définition : Quartiles
Le premier et le troisième quartiles sont des valeurs Q_1 et Q_3 qui séparent l'échantillon en deux parts : .
.
.
.
Remarque : Quartile et médiane
Le deuxième quartile correspond à la médiane.

Méthode de calcul :

1. Calculer la valeur $k = \frac{n+3}{4}$ s'il s'agit du prmeier quartile, ou $k = \frac{3n+1}{4}$ s'il s'agit du troisième quartile.

2.a. Si k est un nombre entier, alors la valeur du quantile correspond à $x_{(k)}$, la $k^{\text{ème}}$ valeur de la série ordonnée.

2.b. Si k n'est pas entier, alors on note i la valeur entière de k et le quartile se calcule avec la formule suivante : .
.

Définition : Ecart inter-quartile (IQR)
Correspond à la distance entre les deux quartiles : $\text{IQR} = Q_3 - Q_1$.

Indicateurs de répartition (1)

Définition : Quantile
.
.
Définition : Quantile empirique
.
.
.
.
Remarque : Quantile et quartile
Les quartiles Q_1 et Q_3 correspond aux quantiles $q_{25\%}$ et $q_{75\%}$, et la médiane au quantile $q_{50\%}$.

Indicateurs de répartition (2)

Définition : Bornes
Les bornes des données correspondent à la valeur minimale et à la valeur maximale observée. On note $b_{\min}$ et $b_{\max}$ ces bornes :
- $b_{\min} = x_{(1)}$, et
- $b_{\max} = x_{(n)}$.

Définition : Amplitude ou étendue
Correspond à l'écart entre les bornes des données.

Résumés graphiques (3)

Boxplot :
Graphique permettant de visualiser grossièrement la distribution des données.

- La première délimitation à gauche du graphique correspond à une borne minimale estimée par la formule suivante, et notée $b_{\min}$:

$$b_{\min} = \max(x_{(1)}, Q_1 - 1.5 \times \text{IQR}).$$

- La seconde délimitation correspond au premier quartile Q_1.
- La troisième délimitation correspond à la médiane M.
- La quatrième délimitation correspond au troisième quartile Q_3.
- La cinquième délimitation correspond à une borne maximale estimée par la formule suivante, et notée $b_{\max}$:

$$b_{\max} = \min(x_{(n)}, Q_3 + 1.5 \times \text{IQR}).$$

En résumé

Un résumé rapide de ce chapitre :

— Type de données
  Distinguer parmi les quantitatives
— Catégorisation et histogramme
— Densité de probabilité
— Indicateurs de position (tendance centrale)
  Moyenne, médiane, mode
— Indicateur de dispersion et théorème de Köning-Huygens
  Variances empiriques, écart-type, coefficient de variation
— Indicateurs basés sur les quantiles
  Mediane, quartiles, IQR, boxplot

TD 2 – Résumés statistiques et représentations graphiques

Exercice 1. On s'intéresse aux mots de cette feuille de TD. En particulier, on va se concentrer sur le nombre de fautes d'ortographe.

1. Comptait le nombre de fotes d'orthographe sur cette feuille. Rassemblez les résultats au tableau et notez de quelle rangée viennent chacune des données.
2. Calculez la moyenne et la variance de ces deux variables.
3. Tracez un boxplot.
4. Tracez un boxplot par rangées.
5. Selon vous, il y a-t-il une différence entre les rangées en terme de données ?

Exercice 2. Le tableau suivant donne la répartition d'une population par tranche d'âge.

Classes d'âge	$[0, 10[$	$[10, 20[$	$[20, 30[$	$[30, 40[$	$[40, 50[$	$[50, 0[$	$[60, 70[$	$[70, 80[$
Nombres	18	44	68	54	42	36	16	10

Calculez la médiane les quartiles de cette série statistique.

Exercice 3. Lors d'un banquet de marriage, on a relevé les âges des 47 convives présents et on a obtenu les valeurs suivante :

$$10, 13, 23, 29, 92, 71, 45, 54, 24, 30, 17, 85, 31, 17, 44, 50, 48, 22, 14, 70, 53, 5, 48, 49, 24, 45, 52, 46, 15, 87, 49, 50,$$
$$29, 51, 23, 24, 53, 60 \text{ et } 65.$$

Représentez cette série de valeurs par un boxplot. Pour vous aider, vous commencerez par classer les 47 valeurs par ordre croissant.

Exercice 4. Une enquête sur les salers mensuels des techniciens de quatre entreprises de la même région et du même secteur a données les résultats suivant (en euros arrondis au cinquante le plus proche) :

Entreprise	Effectif	$x_{\min}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$x_{\max}$
A	110	1050	1500	1750	2000	2350
B	88	1150	1750	2200	2400	2950
C	81	1150	1500	1650	1900	2750
D	25	950	1200	1350	1500	2250

Comparez ces entreprises à l'aide de boxplots tenant compte de leurs effectifs.

Exercice 5. Voici ci-dessous des scores de sommeil (entre 0 et 100) pour 30 individus.

68.75	73.02	81.84
70.37	70.78	81.56
68.33	68.76	80.15
73.19	65.57	76.02
70.66	72.25	81.24
68.36	69.91	79.89
70.97	69.97	79.69
71.48	71.89	77.06
71.15	71.64	79.04
69.39	71.19	80.84

1. Calculez un histogramme de ces données pour les intervalles $[65, 67.5[, [67.5, 70[, \ldots, [77.5, 80]$.
2. Même question pour les intervalles $[65, 70[, [70, 75[, [75, 80[, [80, 85]$.

Chap. 5 – **Données atypiques**

Introduction : les données atypiques

Contexte : Données atypiques
Présentes dans quasiment toutes les bases de données. Souvent indétectable pour un humain. D'autant plus présentes et indétectables lorsqu'il y a beaucoup de données.

Problème : Ne pas chercher à les détecter/étudier peut mener à des résultats erronés.

Problème complexe sans solution uniformément meilleure dans toutes les situations. Les méthodes à utiliser dépendent largement du contexte, des propriétés des données et de l'objectif de l'étude.

Les données atypiques

Définition : Donnée atypique
.
.

Remarque :
Cette dénomination englobe plusieurs notions (non-disjointes).

Définition : Donnée extrême
.
.
Exemple : Niveau de rivière

Définition : Donnée anormale ou anomalie
.
.

Définition : Donnée aberrante
.
.
Exemple : Capteur de température, taille d'une population contenant un joueur de basketball.

Remarque :
Ces notions et appellations sont souvent emmêlés, voire indistinguées. Dans le cadre de ce cours, on fera attention à les distinguer.

Conséquences

Propriété :
.
.

⋆ Déviation de l'estimation : (estimateurs non-robustes)
Salaire des employés dans une entreprise de 23 personnes.

	Statut	Salaire
1	Tech	1358.37
2	Tech	1317.76
3	Tech	1313.27
4	Ing	1879.94
5	Dir	9760.00
6	Tech	1401.77
7	Tech	1298.47
8	Tech	1340.13
9	Cad	2279.57
10	Ing	1854.05
11	Tech	1308.01
12	Tech	1288.63
13	Tech	1360.30
14	Ing	1916.50
15	Tech	1328.34
16	Cad	2221.00
17	Tech	1323.82
18	Tech	1339.26
19	Tech	1310.17
20	Tech	1360.44
21	Tech	1338.80
22	Ing	1969.60
23	Ing	1961.37

.
.
.
.
.
.

Détection des valeurs atypiques

⋆ Pour une variable discrète : donnée possédant une modalité "trop" rare.
.
.
⋆ La donnée associée à "*plus de 250 grammes d'alcool par jour*" est atypique.
⋆ La donnée associée à "*plus de 250 grammes d'alcool par jour*" et "*moins de 9 grammes de tabac par jour*" est atypique.

Détection des valeurs atypiques

⋆ Pour une variable continue :
.
.

Plusieurs critères peuvent être utilisés :
- seuil a priori ou naturel : température qui n'est pas entre 0 et 40 degrés
Le choix du seuil peut certes s'imposer suivant le contexte, mais il est arbitraire.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Détection multivariée

⋆ Pour une variable continue :
donnée dont la valeur (ou autre chose) dépasse un seuil.

Cas de données non-extrêmes :

Les données atypiques sont celles qui sortent de ce que prédit le modèle.

Dans chacun de ces deux cas, le modèle est donné par la courbe rouge et les points noirs sont ceux qui sont trop éloignés de cette courbe.

Traitement des données atypiques

Les traitements doivent se faire avec une annotation.

⋆ Donnée aberrante :
- suppression de la base de donnée (peu recommandé)
- imputation : remplacer la valeur aberrante par
  la moyenne des données,
  la médiane des données,
  la moyenne des données les plus "proches",

une borne définie par une raison dépendant du contexte ou
la valeur d'un quantile.

⋆ Donnée anormale ou extrême :
- suppression de la base de donnée (peu recommandé)
- séparation des analyse :
étudier/décrire/comprendre les données atypiques,
analyser séparément les données non-atypiques et
déterminer l'impact des données atypiques sur le résultat.

⋆ Autre possibilité :
- Utiliser des estimateurs/méthodes robustes.
(moins sensibles aux données atypiques)
(on aime pas supprimer des données : argent, temps, investissement)

Résumé

⋆ Pas de méthode uniformément meilleure dans toutes les situations.
⋆ Atypique : englobe plusieurs cas.
⋆ Impact sur les résultats statistiques.
⋆ Plusieurs manières pour les détecter.
- dépend du type de variable,
- dépend du nombre de données et de variables, ...
- se base sur des critères arbitraires ou subjectifs.
⋆ Plusieurs traitements possibles.

Take home message :

- Attention, supprimer ou modifier les observations atypiques à un modèle sans justification serait totalement contraire à l'éthique. L'objectif est avant tout de les identifier car ce sont celles, les plus susceptibles d'être la conséquence d'une erreur de mesure, de libellé, ou encore une anomalie, défaillance ou tentative de fraude, d'intrusion, selon le contexte.

- Leur détection et leur compréhension peuvent être des objectifs, suivant le contexte, comme parfois dans une problématique de cyber-sécurité.

TD 3 – Quantiles, valeurs atypiques

Exercice 1. Soit la série statistique suivante correspondant à la mesure d'une variable X :

	1	2	3	4	5	6	7	8	9	10
1	2	9	24	36	42	73	74	86	87	101
2	101	103	126	134	145	159	164	167	191	197
3	204	205	221	227	230	267	276	294	300	324
4	327	328	337	341	352	352	365	367	369	388
5	388	390	393	399	421	421	432	439	475	475

1. Quelle est le type de la variable ?
2. Déterminez le premier et le troisième quartile.
3. Déterminez les quantiles à 10%, 20%, 30%, ... et 90%.
4. Sans faire de calculs, déterminez le quantile à 0% et le quantile à 100%.
5. Tracez les quantiles sur un graphique.
6. Tracez un boxplot de cette série statistique, pour lequel la borne minimale et la borne maximale du boxplot correspondent respectivement au quantile à 10% et au quantile à 90%.

Exercice 2. Lors des élections de 2017 en France, un institut de sondage a publié les intentions de votes suivantes toutes les semaines.

	candidat1	candidat2
Semaine 1	19.50	20.10
Semaine 2	19.60	20.00
Semaine 3	19.70	19.80
Semaine 4	19.40	19.80
Semaine 5	19.60	19.70
Semaine 6	20.10	20.10
Semaine 7	20.00	20.00
Semaine 8	19.70	19.80
Semaine 9	19.60	19.70
Semaine 10	20.00	19.70

1. Tracez un graphique représentant l'évolution des intentions de votes pour chacun des candidats.
2. Suite à la publication de l'institut de sondage à la semaine 10, un journal national a choisi comme Une : "Candidat 1 est pour une première fois bien parti pour remporter l'élection". A l'aide du graphique, relativisez cette Une.
3. Tracez un boxplot des intentions de votes pour le candidat 1 et déterminez si lors de la semaine 10, les intentions de votes pour le candidat 1 est une donnée atypique.
4. Si on détermine (comme précédemment) une donnée atypique en se basant sur le graphique d'un boxplot, quelle devrait être au minimum le pourcentage d'intentions de votes pour que la donnée soit considérée comme atypique ?
5. Tracez l'évolution de la différence d'intentions de votes entre le candidat 1 et le candidat 2.
6. Tracez un boxplot des différences d'intentions de votes entre le candidat 1 et le candidat 2. Déduisez-en si la différence d'intentions de votes lors de la semaine 10 est une donnée atypique.

Exercice 3. On s'intéresse dans cet exercice à des données boursières relatives à certaines entreprises.

1. Pour chacune des variables, déterminez les données atypiques en traçant des boxplots.
2. Tracez les dividendes en fonction de la dette (nuage de points).

3. En vous appuyant sur les résultats de la question 1, indiquez quels points de ce graphique correspondent aux données atypiques par rapport aux dividences. Faites de même pour les données atypiques par rapport à la dette.
4. En considérant les données sous un angle bidimensionnel, donnez votre avis sur quelles données sont atypiques.
5. Expliquez le cas particulier des points pour lesquels $y = 0$.

Dans ce qui suit, on prendra seulement en compte les entreprises qui ont versé des dividendes.

6. Tracez le même nuage de points sans les points relatifs aux entreprises n'ayant pas versé de dividendes.
7. Sur ce dernier graphique, tracez les courbes $\mathcal{C}_1$ d'équation $y = 0.01 + \frac{x}{1500}$ et $\mathcal{C}_2$ d'équation $y = 0.025 + \frac{x}{150}$. En déduire une interprétation sur une modélisation possible de la relation dividende/dette d'une entreprise.
8. Associez chaque donnée à un des deux modèles.
9. Indiquez les entreprises qui vous semblent atypiques.

	Rentabilité des actifs	Rendement des capitaux propres	EBE/CA	Dette/EBE	Dividende
Abivax SA	-0.37 %	-0.23 %			
Albioma	0.06 %	0.1 %	-0.16	8.73 %	0.02 %
ATARI		0.22 %		7.68 %	0 %
Aubay SA	0.13 %	0.18 %	0.02	0.99 %	0.01 %
Auplata SA	-0.12 %	-0.25 %			0 %
Bigben Interactive SA	0.19 %	0.07 %		4.97 %	0 %
Cellectis SA	-0.27 %	-0.37 %	-5.72		0 %
Chargeurs SA	0.08 %	0.11 %	-0.03	2.43 %	0.04 %
Claranova SA	-0.08 %	-6.8 %	0.03		0 %
DBV Technologies SA	-1.04 %	-0.8 %			
Devoteam SA	0.15 %	0.16 %	0.04	1.05 %	0.01 %
ERYTech Pharma SA	-0.26 %	-0.31 %			
Genfit SA	-0.29 %	-0.47 %	-947.39		0 %
Getlink SE	0.05 %	0.06 %	-0.02	10.4 %	0 %
Groupe Guillin SA	0.09 %	0.16 %	0.01	1.58 %	0.03 %
Groupe Open SA	0.09 %	0.09 %	0.03	0.47 %	0.02 %
Infotel SA	0.13 %	0.21 %	0.04	0 %	0.03 %
Innate Pharma SA Class A	-0.02 %	-0.56 %	-0.4		0 %
Interparfums	0.11 %	0.1 %	0.05	1.98 %	0.01 %
JACQUET Metal Service SA	0.09 %	0.15 %	-0.01	40.22 %	0.04 %
Kaufman & Broad SA	0.1 %	0.36 %	0.07	1.6 %	0.22 %
Latecoere SA	0.04 %	0.01 %	-0.09	5.07 %	0 %
Lectra SA	0.11 %	0.21 %	0.04	0 %	0 %
LNA Sante SA	0.05 %	0.14 %	0.06	9.35 %	0 %
Lumibird SA	0.08 %	0.06 %	0.01	159.56 %	0 %
Manitou BF SA	0.11 %	0.12 %	-0.04	2.06 %	0.03 %
Nanobiotix SA	-0.65 %	-0.86 %	-227.62		
Nicox SA	-0.15 %	-0.07 %	-4.11		0 %
Pharnext SA	-0.88 %		-6.75		
Plastiques du Val de Loire SA	0.09 %	0.2 %	-0.03	7.5 %	0.04 %
Poxel SA	0.1 %	-0.76 %	-0.06		
Quantum Genomics					
Robertet SA	0.11 %	0.14 %	0.04	2.04 %	0.02 %
Soitec SA	0.17 %	0.43 %	-0.09	19.93 %	0 %
Solocal Group	0.16 %		0.01	8.09 %	0.07 %
Solutions 30 SE		0.26 %		3.4 %	0 %
SRP Groupe SA	-0.02 %	-0.03 %	0	0.34 %	
Trigano SA	0.17 %	0.24 %	0.04	0.51 %	0.02 %
Virbac SA	0.06 %	-0.01 %	0.05	19.23 %	0.02 %
Xilam Animation SA	0 %	0.3 %	-0.22	16.51 %	0 %

CHAPITRE 3

Statistique bivariée

Ce chapitre va contenir une deuxième partie.