Chapitre 1 – Un peu de vocabulaire

Vous avez défini un sujet de recherche, sélectionné des individus dans un ou plusieurs groupes, ou échantillons, déterminé une liste de variables, et noté les valeurs de ces variables sur les individus sélectionnés. Ce livre commence après cette étape, appelée recueil des données, et ne traite pas du choix des données à recueillir, ni du nombre d'individus nécessaires. Son objectif est de vous aider depuis la constitution du fichier de données jusqu'à l'écriture du mémoire ou de l'article, en passant bien sûr par l'analyse statistique des données, qui constitue la partie principale de cet ouvrage. Cependant, avant d'entrer dans le vif du sujet, il est indispensable de définir quelques termes, qui seront utilisés tout au long de ce livre.

Données, individus, variables

Vous avez recueilli des données. Plus précisément, vous avez recueilli les valeurs de certaines *variables* sur des *individus*. Un *individu*, tel qu'il est défini dans ce livre, peut être une personne (homme ou femme, adulte ou enfant) ou un animal (rat, souris...). La *variable* est la caractéristique qu'on mesure ou dont on recueille la valeur sur un individu (l'âge, le sexe, la maladie, les résultats de diverses épreuves...).

Nombre de mesures ou occasions

Le plus souvent, une variable n'est mesurée qu'une seule fois pour chaque individu, mais il arrive qu'on mesure plusieurs fois la même variable. Ceci peut se produire principalement dans trois circonstances :

- quand on veut évaluer l'effet d'un traitement ou d'un apprentissage en mesurant une variable avant et après son administration. On a alors deux mesures par individu. En terme statistique, on est dans le cadre des *séries appariées.*
- Quand on mesure la valeur d'une variable plusieurs fois au cours du temps. Ce peut être, en médecine, pour étudier l'évolution de la concentration d'une molécule ou l'effet d'un traitement, en épidémiologie pour déterminer l'évolution d'une maladie, ou en biologie pour étudier l'effet du vieillissement, naturel ou modifié par des expériences. Statistiquement, on est dans le domaine des *comparaisons de courbes*.
- Enfin, les mesures peuvent être recueillies plusieurs fois sur le même individu par le même observateur ou par des observateurs différents afin d'évaluer respectivement les *concordances intra-juge* ou *inter-juges*.

Variables quantitatives, qualitatives nominales ou ordinales, modalités

Les variables sont de différents types. On distingue :

- Les variables *quantitatives*. Elles ont pour valeurs des nombres (entiers ou décimaux), comme par exemple le nombre d'enfants dans une famille ou l'âge d'un individu.
- Les variables *qualitatives nominales*. Elles ne sont pas quantitatives, et leurs valeurs ne sont pas ordonnées. On peut citer par exemple le sexe d'un individu, le caractère malade ou non, un traitement médicamenteux.

- Les variables *qualitatives ordinales*, ou plus simplement *ordinales*. Ce sont des variables qualitatives dont les valeurs sont naturellement ordonnées. Dans certains questionnaires, par exemple, les réponses à des questions de la forme « Vous arrive-t-il de ressentir telle impression ? » sont : « jamais », « quelquefois », « souvent », toujours ». On peut aussi construire des variables ordinales en regroupant les valeurs d'une variable quantitative en classes. Par exemple, on peut regrouper des âges en « < 20 ans », « 20 – 50 ans », « 50 – 65 ans » et « > 65 ans ». La variable « classe d'âge » ainsi définie est ordinale. On voit sur cet exemple qu'une variable quantitative par nature, comme l'âge, peut être codée comme une variable qualitative ordinale. C'est bien entendu le codage qui compte pour choisir les tests statistiques.

Les valeurs d'une variable qualitative, nominale ou ordinale, s'appellent des *modalités*. Une variable qui a deux modalités, comme le sexe par exemple, est dite « *binaire* », et est en général considérée comme une variable nominale.

Echantillons ou groupes, populations

Comment avez-vous sélectionné les individus sur lesquels vous avez recueilli les données ? Les statistiques sont basées sur le modèle de l'urne : on suppose qu'il y a une ou plusieurs grandes urnes, contenant chacune un nombre très grand de boules possédant des caractéristiques diverses (ces boules sont les individus), et que vous avez tiré au sort des boules dans ces urnes. Chaque urne représente une *population*. Dans chaque urne, vous sélectionnez un certain nombre de boules, dont vous observez ou mesurez des caractéristiques : l'ensemble de ces boules constitue un *échantillon*. Nous emploierons parfois dans ce livre le terme de « *groupe* », plutôt que le terme d'*échantillon*, plus technique. L'objectif de l'étude est de connaître les caractéristiques de l'ensemble des boules que contiennent les urnes, qu'on ne pourra jamais observer toutes, mais dont on peut observer un *échantillon*. Les ensembles des boules contenues dans les urnes constituent les *populations*. La *statistique* est la science qui permet de déduire (le vrai terme est « inférer ») les caractéristiques des populations à partir des échantillons. Il est important, pour interpréter les résultats, d'avoir une idée des populations concernées par l'étude.

Statistique descriptive ou inférentielle

Une partie des statistiques vise à résumer l'information contenue dans un jeu de données (un échantillon). Cette partie est la *statistique descriptive*. Une autre partie a pour but d'inférer à partir des données observées des connaissances concernant les populations d'où on a tiré au sort les échantillons. Cette partie des statistiques constitue la statistique inférentielle. Elle est elle-même divisée en deux parties : estimation et tests.

Estimation

L'estimation a pour objectif de connaître la valeur de certaines caractéristiques de la population. On cherchera par exemple à déterminer la probabilité pour un individu atteint d'une maladie d'être guéri par un traitement. Cette probabilité, d'après le modèle de l'urne, est numériquement égale à la proportion d'individus guéris par le traitement dans la population des individus atteints de la maladie. Ces deux quantités vont être estimées par la proportion d'individus guéris dans un échantillon d'individus malades.

Paramètres, estimateurs

Les grandeurs dont on cherche à déterminer la valeur dans la population sont appelées des *paramètres*. Les paramètres les plus fréquemment utilisés en statistique sont la *proportion*, la *moyenne* et l'*écart-type*. On accole en général à ces paramètres l'adjectif « vrai » pour indiquer qu'on s'intéresse à leur valeur dans la population. On parle donc de *proportion vraie*, de *moyenne vraie*, et d'*écart-type vrai*. Ces paramètres sont estimés à l'aide des données observées sur un échantillon par des *proportions*, des *moyennes* et des *écarts-types* calculés à partir des valeurs observées. Ces valeurs calculées sont dites *observées*, ou *expérimentales*, pour indiquer qu'elles concernent l'*échantillon* et non la *population*. Les *proportions observées*, *moyennes observées* et *écarts-types observés* sont des *estimateurs* des *proportions vraies*, *moyennes vraies* et *écarts-types vrais*, respectivement.

Intervalles de confiance

Les caractéristiques des populations (*proportions*, *moyennes*,…) ne peuvent pas être connues exactement. C'est pourquoi les résultats des estimations sont données en général sous la forme d'un intervalle, dit « *intervalle de confiance* ». C'est un intervalle qui contient la valeur qu'on cherche à estimer avec une probabilité donnée, en général de 95 %. On parle d'intervalles de confiance de niveau 95 %.

Tests

Les tests statistiques servent à démontrer scientifiquement des propositions. Par exemple, on cherchera à démontrer qu'un traitement A est meilleur qu'un traitement B pour guérir une maladie donnée. Pour cela, il faut formuler le problème en termes statistiques. On n'entrera pas dans ce livre dans les formulations techniques, mais on peut dire qu'on utilisera un *test* qui permettra de démontrer, ou pas, que la probabilité de guérison du traitement A est supérieure à celle du traitement B. Cette conclusion sera basée sur les proportions de guérisons observées sur les échantillons, et ne peut être formulée dans le sens désiré que si les proportions observées sont déjà dans le bon sens : on n'a bien sûr aucune chance de démontrer que le traitement A est meilleur si sa proportion observée de guérisons est la plus faible. Si on a observé des proportions qui diffèrent dans le bon sens, la question est de savoir si cette différence est due au hasard ou non. C'est à cette question que répondent les *tests*, en déterminant une valeur à partir de laquelle on peut conclure que la différence observée n'est vraisemblablement pas due au hasard, et donc qu'il existe une différence de proportions dans les populations.

Risques, puissance

La *statistique inférentielle* conclut à propos de *populations* qu'on n'observe jamais en totalité. Il y a donc toujours un risque pour que les conclusions énoncées soient fausses. On cherche en général à démontrer des différences, et les risques pris sont en réalité de nature différente :

- On peut conclure à une différence qui n'existe pas. Ce type d'erreur est contrôlé par les statisticiens, et fixé à 5 %. On appelle souvent ce risque le *risque α*.
- On peut ne pas conclure à une différence alors qu'en réalité il en existe une. Ceci survient d'autant plus que la vraie différence est faible et que les effectifs des échantillons sont faibles. On définit la *puissance* d'un test comme sa probabilité de démontrer l'existence d'une différence quand elle existe. Le risque décrit dans ce paragraphe est donc le *manque de puissance*,

noté parfois *risque β*. Il n'est jamais connu parfaitement en pratique, puisqu'on ne connaît pas les vraies différences, mais on peut le calculer en faisant des hypothèses sur les valeurs des *paramètres* des *populations*.

p-value

Quand on énonce une conclusion comme « le traitement A est meilleur que le traitement B », le risque d'erreur de cette phrase est de 5 %. Ceci veut dire plus précisément : « En l'absence de différences dans les populations, il y avait moins de 5 chances sur cent d'observer la différence effectivement observée ». On accompagne l'énoncé de la conclusion d'une *p-value*, qui est, en première approximation, la probabilité d'observer une différence aussi grande que celle qu'on a effectivement observé, s'il n'y a aucune différence dans les populations. Cette *p-value* est bien entendu inférieure à 5 %, puisqu'on a trouvé une différence, mais sa valeur sert à contraster les conclusions probables, avec des *p-values* proches de 0,05, et les conclusions quasi-certaines, avec par exemple des *p-values* inférieures à 0,000001, qui signifient qu'il y a moins de une chance sur un million pour que la différence observée ne soit due qu'au hasard.

Chapitre 2 – Constituer un fichier de données « propre »

Vous avez recueilli des données, sans doute en cochant ou en écrivant des réponses sur des questionnaires, à la main, sur du papier. Il faut maintenant les entrer dans un fichier informatique. Ce chapitre vise à vous aider à élaborer un fichier de données « propre », c'est-à-dire lisible par un logiciel de statistique. Les exemples de ce livre seront réalisés avec le logiciel XLSTAT, mais tous les logiciels de statistiques présentent de grandes similarités, et les notions de ce chapitre, plus particulièrement, s'appliquent à tous les logiciels.

La manière de représenter les informations de manière à ce qu'elles soient « compréhensibles » par un ordinateur s'appelle le *codage*. Ce chapitre traitera donc de la présentation générale des *données* (nombre et structure des fichiers) et du *codage* des noms des *variables*, des *individus*, puis des valeurs des variables.

Pour illustrer ce chapitre, on utilisera un sous-ensemble de données d'une étude qui vise à créer un test de performances concernant la lecture chez des élèves du collège. Dans trois écoles, et dans toutes les classes de la sixième à la troisième, des élèves ont passé des tests de lecture. En tout on a fait passer les épreuves à 144 élèves. On a recueilli quelques informations générales sur chaque élève, telles que son sexe, son âge, s'il aimait lire, s'il était un lecteur régulier, puis on lui a fait passer une série d'épreuves. Parmi celles-ci, on a noté les résultats d'une épreuve consistant à lire 5 mots (on note le nombre de mots bien lus), le temps mis pour lire un texte, et enfin une appréciation de la compréhension du texte par l'élève.

Une partie des données recueillies se présente sous la forme de la figure 2.1. Cette figure représente une partie d'une feuille EXCEL contenant l'ensemble des données recueillies. La présentation adoptée est « naturelle », c'est-à-dire compréhensible par un être humain, mais très loin d'une forme « lisible » par un logiciel de statistique. Nous allons dans ce chapitre détailler les étapes qui transformeront les données initiales en une feuille EXCEL lisible par l'ordinateur.

1. Il faut un seul tableau autant que possible

Dans la feuille EXCEL initiale (figure 2.1), il y a 3 tableaux : (Ecole A, classe de 3^{e}), (Ecole A, classe de 4^{e}), et (Ecole B, classe de 3^{e}). Il faut faire tenir l'ensemble des données dans un seul tableau. C'est toujours possible quand chaque variable n'est mesurée qu'une fois pour chaque individu : il faut constituer un tableau avec une ligne par individu et une colonne par variable. Ici, le nom de l'école et la classe doivent être considérés comme deux variables en plus. En ajoutant les deux variables Ecole et Classe, et en « collant » les trois tableaux en un seul, on obtient le tableau intermédiaire de la figure 2.2.

Figure 2.1 : données initiales

	A	B	C	D	E	F	G	H	I
1	**Ecole A**			**Classe 3e**			**tests**		
3	**Nom**	**Prénom**	**Sexe**	**Âge**	**A l'habitude de lire**	**Aime lire**	**Lecture de 5 mots**	**Temps de lecture**	**Compréhension**
4	martin	emma	F	14 ans 6 mois	oui	passioné-ment	4 mots	119 sec	oui
5	bernard	clara	F	14 ans 8 mois	oui	pas du tout	4 mots	79 sec	bonne
6	dubois	lucas	G	15 ans 3 mois	non	beaucoup	5 mots	107 secondes	O
7									
8	**Ecole A**			**Classe 4e**					
10	**Nom**	**Prénom**	**Sexe**	**Âge**	**A l'habitude de lire**	**Aime lire**	**Lecture de 5 mots**	**Temps de lecture**	**Compréhension**
11	thomas	enzo	M	13 ans 1 mois	Oui	un peu	5	1 mn 15 s	Oui
12	robert	lola	F	13 ans 4 mois	O	un peu	5	1 mn 30 s	Non
13	richard	nathan	M	13 ans 9 mois	N	pas du tout	5	1 mn 27 s	Non
14									
15	**Ecole B**			**Classe 3e**					
17	**Nom**	**Prénom**	**Sexe**	**Âge**	**A l'habitude de lire**	**Aime lire**	**Lecture de 5 mots**	**Temps de lecture**	**Compréhension**
18	petit	thomas	G	15 ans 11 mois	N	pas du tout	5	88	o
19	durand	manon	F	??	?	pas du tout	5	120	o
20	leroy	maxime	G	14 ans 7 mois	ne sait pas	un peu	4	102	n

Figure 2.2 : tableau intermédiaire

	A	B	C	D	E	F	G	H	I	H	I
1											
3											
4									**Tests**		
5	**Ecole**	**Classe**	**Nom**	**Prénom**	**Sexe**	**Âge**	**A l'habitude de lire**	**Aime lire**	**Lecture de 5 mots**	**Temps de lecture**	**Compréhension**
6	A	3	martin	emma	F	14 ans 6 mois	oui	passionément	4 mots	119 sec	oui
11	A	3	bernard	clara	F	14 ans 8 mois	oui	pas du tout	4 mots	79 sec	bonne
12	A	3	dubois	lucas	G	15 ans 3 mois	non	beaucoup	5 mots	107 secondes	O
13	A	4	thomas	enzo	G	13 ans 1 mois	Oui	un peu	5	1 mn 15 s	Oui
18	A	4	robert	lola	F	13 ans 4 mois	O	un peu	5	1 mn 30 s	Non
19	A	4	richard	nathan	G	13 ans 9 mois	N	pas du tout	5	1 mn 27 s	Non
20	B	3	petit	thomas	G	15 ans 11 mois	N	pas du tout	5	88	o
21	B	3	durand	manon	F	??	?	pas du tout	5	120	o
22	B	3	leroy	maxime	G	14 ans 7 mois	ne sait pas	un peu	4	102	n

Dans le tableau intermédiaire de la figure 2.2, il faut :

2. Enlever les noms et prénoms

Les noms et prénoms des élèves figurent dans ce tableau. (J'ai choisi les noms de famille et les prénoms d'après leur ordre de fréquence en France en 2009 ; toute homonymie ne serait que le résultat d'une coïncidence). Il faut supprimer du tableau les colonnes Noms et Prénoms, et les remplacer par un numéro. On va donc remplacer les variables « Nom » et « Prénom » par la variable « NumEnfant », qui prendra les valeurs de 1 à 144. On peut conserver dans un autre fichier la correspondance entre nom, prénom et numéro.

3. Supprimer les lignes vides

Il y a deux lignes vides avant le début du fichier. Il faut les supprimer : la première ligne du fichier ne doit contenir que les noms des variables.

4. Pour le codage des noms de variables : une cellule par nom

Les noms de *variables* doivent être écrits dans une seule cellule (case). Ici, le nom de la variable « Tests de lecture de 5 mots » est écrit dans deux cases. Il faut l'écrire dans une seule case.

5. Codage des noms de variables

Les noms des *variables* sont longs, constitués de plusieurs mots, et contiennent des accents. Il faut recoder les noms des variables en un seul mot, sans blanc au milieu du nom, sans accent ni caractères spéciaux (comme « > » ou « < ») ni signe de ponctuation. Les noms de doivent contenir que des lettres et des chiffres. Ils doivent commencer par une lettre. Enfin, ils ne doivent pas être trop longs : il faut enlever les articles, abréger les mots. Par exemple, le « Tests de lecture de 5 mots » peut devenir « TestLecture5 ». (Quand le nom de la variable est constitué de plusieurs mots, on peut commencer chaque ancien mot par une majuscule, de manière à améliorer la lisibilité). Enfin, bien sûr, il faut que chaque nom de variable ne figure qu'une fois dans le fichier.

6. Codage des variables quantitatives

La règle est simple : une *variable quantitative* doit être représentée sous la forme d'un nombre par cellule, sans aucune lettre dans la même cellule, ni dans aucune cellule de la même colonne. Les âges du tableau intermédiaire seront considérés par un logiciel comme des variables qualitatives, car ils comportent des lettres (« ans », « mois »). Aucun logiciel ne sait manipuler des âges codés de cette manière. La même considération s'applique aux temps de lecture exprimés sous la forme « 1 mn 15 s », ou « 79 s ». Seuls les temps des 3 dernières lignes (88, 120, 102) seront reconnus et traités par un logiciel de statistique.

7. Cas particulier : codage des durées et des âges

- Pour les durées longues, telles que des âges ou des durées de survie, l'unité de base doit être le jour. Il faut entrer dans votre questionnaire, comme données initiales, les dates, exprimées dans le format jj/mm/aaaa (comme par exemple noël 2010 sera le 25/12/2010). Pour coder une durée entre deux événements ou un âge à une date donnée en jours, il faut créer une colonne pour cette nouvelle variable, puis inscrire dans la première cellule une formule correspondant à la différence (date de l'examen – date de naissance), comme par exemple « age en jour = date examen – date de naissance », puis copier-coller cette formule pour toute la colonne. Si EXCEL vous montre cette durée sous la forme d'une date, pas de panique,