Chapitre 1: Variables et niveaux de mesure

Ce premier chapitre sera consacré à quelques notions fondamentales dont la compréhension sera nécessaire à la bonne application des différents types d'analyse que nous aborderons dans tous les chapitres ultérieurs de cet ouvrage. Il s'agit donc d'un chapitre essentiel. Mon expérience personnelle d'enseignant en statistique et de conseiller auprès de chercheurs m'a appris que beaucoup de difficultés majeures lors des examens (pour les étudiants) et lors de l'analyse des données d'une étude (pour les chercheurs) proviennent d'une mauvaise intégration des concepts que nous développons dans le chapitre 1. C'est pourquoi, nous terminerons par un point spécifique sur les erreurs typiques dans l'utilisation des concepts fondamentaux ainsi que sur leurs conséquences (presque toujours dramatiques pour les conclusions tirées à partir des données).

1.1 Le concept de variable

Toute étude scientifique, que ce soit en psychologie, en neurosciences ou dans n'importe quelle autre discipline, commence généralement par une phase d'observation ou d'expérimentation. Dans un second temps, les observations et mesures effectuées seront résumées, analysées, le plus souvent par des traitements statistiques, puis interprétées. Lors de cette phase d'observation, il est possible d'identifier des unités d'observations. Une **unité d'observation** est l'élément du réel sur lequel porte l'observation. Selon la discipline scientifique, les unités d'observation peuvent être des sujets humains, des animaux, des plantes, des événements, des objets, des sociétés, ou toute autre chose pouvant être observée. En psychologie, les unités d'observations sont habituellement des sujets humains ou animaux. En neurosciences, il s'agit aussi souvent de sujets humains ou animaux, bien que d'autres types d'unités d'observations comme par exemple des cellules, des tranches de cerveau ou des échantillons de tissus, soient aussi fréquents. Dans cet ouvrage, nous utiliserons souvent le terme générique de « **sujet** » pour qualifier les unités d'observation qui serviront de base aux analyses statistiques. La majorité des exemples proposés sont en effet des études qui portent sur des sujets humains ou animaux. Il faudra toutefois garder à l'esprit que d'autres types d'unité d'observation ne sont pas rares et que les analyses présentées peuvent s'appliquer de manière identique quelle que soit l'unité

d'observation sélectionnée. En règle générale, une étude scientifique comprend plusieurs unités d'observations d'un même type (plusieurs sujets), à l'exception notable d'un type particulier d'étude que l'on nomme « l'étude de cas » qui porte sur une seule unité d'observation, habituellement un sujet.

Toute la statistique repose sur un concept de base, celui de **variable**. *Une variable est une propriété d'une unité d'observation (d'un sujet) qui peut prendre différentes valeurs*. Comme nous le verrons bientôt, il ne s'agit pas nécessairement de valeurs chiffrées. La préférence manuelle par exemple est une variable dont les trois valeurs possibles sont « Gaucher », « Droitier » et « Ambidextre ». La préférence manuelle est donc incontestablement la propriété d'un sujet qui peut prendre différentes valeurs.

Différentes terminologies sont utilisées pour qualifier la valeur qu'obtient un sujet particulier sur une variable. Lorsque les valeurs que peut prendre une variable sont des catégories ou des étiquettes, on parlera plutôt des différentes **modalités** de la variable. Ce sera habituellement le cas pour toutes les variables nominales (que nous définirons au point 1.2). Par exemple, la variable « préférence manuelle » a trois modalités : gaucher, droitier et ambidextre. Pour un sujet particulier, on dira par exemple qu'il se caractérise par la valeur ou la modalité « gaucher » sur la variable « préférence manuelle ». Lorsque les valeurs d'une variable sont chiffrées, on parle plus couramment du **score** d'un sujet sur cette variable. Par exemple, si on demande à un sujet de remplir une échelle chiffrée de mesure de la dépression, on parlera du score de dépression de ce sujet. Le terme « score » sera ainsi de préférence utilisé pour qualifier la valeur d'un sujet sur une variable métrique ou ordinale (voir point 1.2). Notons toutefois que ces règles d'utilisation des termes « modalité » et « score » sont loin d'être universelles. Ces deux termes sont parfois utilisés indistinctement pour qualifier la valeur qu'obtient un sujet, que la variable d'intérêt soit chiffrée ou non.

Pour être valablement traitée avec les outils statistiques que nous aborderons tout au long de cet ouvrage, *une variable doit impérativement affecter à chaque sujet ou unité d'observation une et une seule valeur*. En d'autres termes, chaque sujet obtient un seul score ou une seule modalité sur une variable particulière. S'il y a plusieurs valeurs par individu, il faut nécessairement créer plusieurs variables ou redéfinir plus précisément la variable d'intérêt. Par exemple, si on interroge des sujets sur leur profession, certains d'entre eux risquent de donner plusieurs réponses. Il n'est en effet pas rare qu'une même personne exerce simultanément

plusieurs professions. Pour traiter cette variable, il sera nécessaire de la redéfinir. Par exemple, on pourrait poser la question : « Quelle est votre profession principale ? », quitte à poser une seconde question sur la profession secondaire et à définir une seconde variable si l'information est utile pour l'étude.

Avant de poursuivre, quelques mots de notation sont nécessaires à la bonne compréhension de la suite de cet ouvrage. En règle général, une variable est symbolisée par une lettre majuscule, souvent X ou Y. La variable, et donc la lettre qui la symbolise, servent à définir une propriété qui varie d'un sujet à l'autre (d'une unité d'observation à l'autre). Par exemple :

A = « âge », pour définir la variable qui encode l'âge des sujets.
X = « genre », pour définir la variable qui encode le genre des sujets de l'étude.
Y = « score de dépression », pour définir la variable qui encode le score obtenu par les différents sujets sur une échelle de dépression.

Une valeur particulière (celle d'un sujet) est représentée par la lettre majuscule définissant la variable, suivie d'un indice. Par exemple pour la variable A définie ci-dessus, si Jean (qui a 25 ans) est le huitième sujet d'une étude, nous écrirons :
$A_8 = 25$ (L'indice indique que l'on prend le score du 8ème sujet).

Les lettres majuscules avec indice sont très souvent utilisées dans les formules mathématiques. Par exemple, on rencontrera souvent des formules de ce type :

$$\sum_{i=1}^{5} X_i$$

Elle signifie additionner (le symbole Σ) les scores des sujets sur la variable X à partir du sujet 1 (i = 1) jusqu'au sujet 5. Si les cinq premiers sujets ont les scores : 5, 8, 12, 7 et 2, alors :

$$\sum_{i=1}^{5} X_i = 5 + 8 + 12 + 7 + 2 = 34$$

En statistiques, comme on utilise le symbole « N » pour désigner le nombre total de sujets d'une étude, on rencontre très souvent la formule suivante :

$$\sum_{i=1}^{N} X_i$$

Elle signifie additionner tous les scores de la variable X du premier au dernier sujet de l'étude. La plupart du temps, cette expression est néanmoins simplifiée en : ΣX qui indique simplement l'addition de tous les scores obtenus dans une étude pour la variable X.

1.2 Les niveaux de mesures

Il existe trois types de variables. On parle aussi de trois niveaux d'échelle de mesure. Leur distinction est très importante. Comme nous le verrons au long de cet ouvrage les différents types de variables impliquent des traitements statistiques différents et ne permettent pas de répondre aux mêmes questions. Une mauvaise identification du type de variable en jeu dans une étude est souvent à la source d'erreurs graves dans le traitement des données et aboutit presque immanquablement à des conclusions erronées. Pour effectuer correctement des analyses statistiques, il est donc essentiel de bien intégrer la distinction entre les différents types de variables.

Variables nominales (étiquettes)

Le niveau le plus simple de mesure d'une variable consiste à noter la catégorie à laquelle appartient un sujet. Pour chaque sujet d'une étude, on peut noter par exemple la couleur de ses yeux, son genre, sa profession, son état civil (célibataire, marié, veuf, divorcé), sa religion, etc… Même s'il ne s'agit pas de mesure au sens où on l'entend habituellement, c'est-à-dire de mesure chiffrée, tous ces exemples sont bel et bien des variables qui répondent à la définition fournie ci-dessus : une variable est une propriété d'un sujet qui peut prendre différentes valeurs. Dans le cas de l'état civil par exemple, il s'agit bien d'une propriété d'un sujet qui peut prendre les valeurs « célibataire », « marié », « veuf » ou « divorcé ». Les modalités d'une variable nominale correspondent uniquement à des noms (d'où l'expression variable nominale), des catégories ou des étiquettes.

Les variables nominales n'impliquent jamais de valeurs chiffrées. Elles ne fournissent donc pas d'informations qui permettent de comparer quantitativement des sujets. Chaque sujet se voit simplement attribuer une étiquette. On dira par exemple que tel sujet est de genre masculin, est divorcé, est un employé... Pour bien distinguer une variable nominale des autres catégories de

variables qui seront bientôt décrites, il faut s'interroger sur l'information fournie par la comparaison de deux sujets. Une variable nominale permet uniquement de dire si deux sujets ont des valeurs semblables ou différentes. Elle ne permet pas de classer les sujets les uns par rapport aux autres en fonction d'un ordre logique ou naturel. Pour la variable état civil par exemple, on peut dire qu'un sujet « marié » est différent d'un sujet « veuf », mais on ne peut pas dire qu'il présente un état civil supérieur ou inférieur à celui-ci. Les valeurs d'une variable nominale sont donc des catégories sans ordre.

Deux remarques s'imposent afin d'éviter toute confusion quant à la qualité non chiffrée des variables nominales :

1. Il arrive fréquemment que des variables nominales soient encodées sous forme de nombres. Un chercheur pourrait par exemple décider d'encoder dans sa base de données pour la variable état civil le chiffre « 1 » à la place de « célibataire », le chiffre « 2 » à la place de « marié », le chiffre « 3 » à la place de « veuf » et le chiffre « 4 » à la place de « divorcé ». Sa base de données ressemblerait alors à ceci :

Tableau 1.1 – Exemple de données avec une variable nominale encodée sous forme chiffrée

Sujet N°	Etat civil	Age
1	1	23
2	3	62
3	2	42
4	2	55
5	4	36
6	1	28
….	…	….

Le fait d'encoder une variable nominale sous forme chiffrée ne modifie en rien ses propriétés. Cette variable reste une variable nominale qui ne permet pas de comparer quantitativement les sujets. On ne pourrait donc pas dire qu'un sujet ayant une valeur de « 3 » (sujet veuf) présente un état civil trois fois supérieur à un sujet ayant une valeur de « 1 » (sujet célibataire). Les nombres ne sont ici que des symboles qui ont été arbitrairement fixés par le chercheur. Il existe aussi des variables nominales qui se présentent naturellement sous forme chiffrée mais qui restent néanmoins des variables nominales ne permettant pas de comparaison quantitative. C'est le cas par exemple des codes postaux ou des préfixes téléphoniques. Il est important de

garder à l'esprit que toutes ces variables restent nominales, et n'autorisent donc pas l'utilisation de traitements statistiques qui exigent des niveaux de mesures supérieurs (ordinal ou métrique).

2. Les variables nominales d'une étude sont très souvent représentées sous la forme d'un tableau de fréquence. Comme nous le verrons dans le chapitre 2, un tableau de fréquence consiste à dénombrer les sujets qui obtiennent chacune des modalités d'une variable et à les présenter sous forme d'un tableau résumé. L'exemple 1.1 montre un tableau de fréquence pour les résultats d'une étude dans laquelle l'état civil de 20 sujets a été noté. On voit dans cet exemple que 10 sujets de l'étude sont mariés, 5 sujets sont célibataires, 3 sujets sont divorcés et 2 sujets sont veufs.

Exemple 1.1 Représentation d'une variable nominale sous forme d'un tableau de fréquence

Variable « état civil », valeurs obtenues par 20 sujets :
Marié, Marié, Veuf, Célibataire, Marié, Célibataire, Divorcé, Marié, Marié, Célibataire, Divorcé, Divorcé, Célibataire, Veuf, Marié, Marié, Marié, Célibataire, Marié, Marié.

Etat civil	Fréquence
Marié	10
Célibataire	5
Divorcé	3
Veuf	2

Cette forme de présentation provoque parfois une certaine confusion quant au type de variable en jeu. Le tableau de fréquence donne parfois l'impression que la variable mesurée est la fréquence, variable qui serait alors chiffrée et permettrait donc des comparaisons quantitatives. De manière erronée, on risque alors d'identifier dans ce tableau une variable métrique et de la traiter comme telle dans les analyses statistiques. En réalité, c'est l'état civil qui est la variable de l'étude et le tableau de fréquence n'est qu'une manière commode de résumer les données de l'étude pour cette variable. Dans notre exemple, les unités d'observations sont les sujets de l'étude et les valeurs que les sujets peuvent obtenir sur la variable état civil sont « célibataire », « marié », « divorcé » et « veuf ». Il s'agit bien d'une variable nominale qui attribue des étiquettes aux sujets.

En psychologie expérimentale et en neurosciences, de nombreuses expériences consistent à soumettre un groupe de sujets à un traitement expérimental particulier et à le comparer ensuite à un autre groupe de sujets qui n'a pas reçu ce traitement. Ce deuxième groupe est qualifié de groupe contrôle ou de groupe témoin. Dans ce type d'étude, le fait d'appartenir au groupe contrôle ou au groupe expérimental est une variable nominale qui sera traitée comme telle dans les analyses statistiques des résultats. Sur la variable « traitement reçu », chaque sujet de l'étude se voit donc attribuer soit la valeur « traitement expérimental », soit la valeur « traitement contrôle ». Il s'agit bien d'une variable nominale puisque les différentes valeurs possibles sont des étiquettes. La distinction groupe expérimental/groupe contrôle est en neurosciences et en psychologie expérimentale l'une des variables nominales les plus fréquentes. Voici quelques exemples d'autres variables nominales que l'on rencontre fréquemment dans des études de psychologie ou de neurosciences : le genre, le diagnostic psychiatrique (exemple : schizophrène, dépressif…), le type de médicament ou de substance pharmacologique administrés.

Variables ordinales

Les variables ordinales classent les sujets selon un ordre sous-jacent et permettent donc une comparaison hiérarchique approximative de deux valeurs. Une variable ordinale est une variable dont les valeurs peuvent être ordonnées, d'où l'expression variable ordinale. Ce sont des variables d'ordre comparatif. Les différents scores peuvent être classés et comparés mais aucune information précise n'est fournie quant à la différence entre deux scores particuliers. Les appréciations sur un diplôme (réussite, satisfaction, distinction, grande distinction) constituent un exemple de variable ordinale. On sait qu'une distinction est supérieure à une satisfaction, mais on ne peut pas dire qu'elle est deux fois supérieure par exemple. Un autre exemple classique de variable ordinale en psychologie est la classe sociale. Beaucoup d'études notent la classe sociale des sujets, bien que la façon d'identifier les classes sociales puisse varier considérablement d'une étude à l'autre. Pour en rester à un exemple très simple, considérons une étude dans laquelle trois classes sociales sont identifiées : classe sociale « inférieure », « moyenne » et « supérieure ». La classe sociale ainsi définie est une variable ordinale. On sait qu'un sujet de la classe moyenne fait partie d'une classe sociale plus élevée qu'un sujet de la classe inférieure, mais on ne peut pas dire que sa classe sociale est deux fois, ou trois fois supérieure. En psychologie, de très nombreuses mesures doivent être classées comme des

variables ordinales plutôt que métriques. C'est le cas par exemple de la plupart des échelles proposées dans les questionnaires (voir encart 1.1).

Toutes les formes de classements constituent aussi des variables ordinales. Un classement consiste à ranger des personnes, des objets ou des événements selon un ordre précis. Par exemple, on peut demander à des sujets de ranger des photos de visages selon leur ordre de préférence. On attribue alors une valeur de « 1 » au premier classé, de « 2 » au deuxième, etc… Les ordres d'arrivée lors d'épreuves sportives sont aussi des classements. C'est le cas par exemple du classement d'arrivée d'un marathon. On dira d'un athlète qu'il était 1er, 2ème ou 3ème, ce qui constitue son score sur un classement. Comme toute variable ordinale, les classements permettent de comparer les sujets, mais ne fournissent pas d'information précise sur la différence entre deux scores. On sait par exemple que le visage classé premier est préféré au visage classé deuxième, mais on ne peut dire s'il y avait une forte différence de préférence entre les deux visages. De la même manière, pour le classement d'arrivée du marathon, on sait que le premier a fait mieux que le deuxième. Mais on ne peut préciser la différence exacte entre les deux athlètes. Le premier pourrait être arrivé une seconde ou cinquante minutes avant le deuxième.

Variables métriques

Une variable métrique assigne un score qui permet une estimation précise de la différence entre deux sujets. Elle se mesure en unités standards qui permettent la réalisation d'opérations arithmétiques telles que l'addition ou la soustraction. Avec une variable métrique, il est non seulement possible de dire si un sujet a obtenu un score supérieur ou inférieur à un autre (ce que permettaient déjà les variables ordinales), mais on peut préciser de combien d'unités ce score est supérieur ou inférieur. Si on prend par exemple la variable « taille des sujets », on sait qu'un sujet mesurant 1m85 a une taille supérieure d'exactement 15 cm par rapport à un sujet mesurant 1m70. En psychologie, une variable métrique souvent utilisée est le temps, que ce soit le temps de réponse en millisecondes pour réagir à un stimulus ou le temps en minutes pour réaliser une certaine tâche. Pour prendre un second exemple, on sait qu'un sujet qui a réalisé une tâche en 15 minutes a pris exactement 5 minutes de plus qu'un sujet qui l'a réalisée en 10 minutes. Pour qu'une variable puisse être qualifiée de métrique, il faut en outre que les unités de mesure soient constantes tout au long de l'échelle. Pour reprendre l'exemple de la taille, une

différence de 5 cm a la même signification entre 1m60 et 1m65 qu'entre 1m90 et 1m95. Cette propriété de constance des unités de mesure est souvent problématique pour de nombreuses échelles créées en psychologie. C'est le cas en particulier pour les échelles de Likert (voir encart 1.1). C'est pourquoi il est souvent plus prudent de considérer ces échelles comme des variables ordinales.

Les variables métriques telles que nous venons de les définir peuvent encore se subdiviser en échelles d'intervalles et échelles de rapport (voir ci-dessous). Toutefois, cette distinction n'est pas essentielle à la réalisation de la majorité des traitements statistiques. Elle est d'autant plus superflue qu'en psychologie de nombreuses variables sont difficiles à classer dans l'une ou l'autre de ces deux catégories.

Pour aller plus loin

Plutôt que de parler de variables métriques, la théorie des échelles de mesure distingue généralement les échelles d'intervalles et les échelles de rapports. Cette distinction importante en théorie l'est beaucoup moins dans la pratique statistique courante. En effet, les variables impliquant des échelles d'intervalles ou de rapports sont généralement traitées avec des analyses statistiques identiques. Apprendre à les distinguer n'est donc pas fondamental pour analyser les données communément récoltées dans les études de neurosciences ou de psychologie.

La distinction entre échelles d'intervalles et de rapports porte essentiellement sur l'existence d'un véritable point zéro sur l'échelle de mesure. Un vrai point zéro signifie un score « 0 » correspondant à l'absence de la chose mesurée.

*Dans **les échelles d'intervalles**, il n'y a pas de vrai point zéro. Lorsqu'un point zéro existe, il s'agit d'un point arbitraire qui ne correspond pas à l'absence de la chose mesurée. Pour illustrer les échelles d'intervalles, on utilise traditionnellement les mesures de température qui constituent une des échelles d'intervalles les plus courantes. Les températures peuvent être mesurées en degrés Celsius ou en degrés Fahrenheit, deux exemples d'échelles d'intervalles. Sur ces deux échelles, le point zéro est arbitraire. On ne peut pas affirmer que 0° Celsius ou Fahrenheit correspond à une absence de température. Pour l'échelle Celsius, le point zéro a été fixé arbitrairement à la température à laquelle l'eau gèle.*

***Les échelles de rapports** au contraire se caractérisent par un véritable point zéro. Comme exemples d'échelles de rapports, on peut citer les échelles physiques habituelles : la longueur, le temps, le poids, etc… Toutes ces mesures se caractérisent bien par un point zéro absolu.*

Ainsi par exemple, une distance de zéro cm correspond bien à l'absence de distance et une durée de zéro seconde correspond effectivement à l'absence de temps.

L'existence d'un véritable point zéro a des implications importantes sur les possibilités de comparaison entre des scores différents. Avec les échelles d'intervalles, une différence entre deux scores a une signification précise qui reste identique quel que soit le point de l'échelle. Ainsi par exemple, il y a la même différence de température entre 10 et 12 °C qu'entre 22 et 24 °C. Dans les deux cas, on peut affirmer qu'il y a précisément une différence de 2 degrés. Par contre, les échelles d'intervalles n'autorisent pas d'effectuer des rapports (qui impliquent des divisions) entre les scores. Par exemple, on ne peut pas affirmer qu'une température de 20 °C correspond à deux fois une température de 10 °C. Si vous doutez de cette dernière affirmation, une comparaison des températures mesurées en degrés Celsius ou Fahrenheit permettra d'éclairer le propos. Le tableau 1.2 rapporte la correspondance entre quelques températures mesurées en degrés Celsius et Fahrenheit.

Tableau 1.2 – Correspondance des températures sur les échelles Celsius et Fahrenheit

Celsius (°C)	*Fahrenheit (°F)*
30	*86*
20	*68*
10	*50*
0	*32*

En examinant le tableau 1.2, on comprend mieux que le calcul de rapports entre températures n'a pas de sens. Par exemple, en degrés Celsius, la température double entre 10 et 20 °C. Les mêmes températures exprimées en degrés Fahrenheit passent de 50 à 68 °F, ce qui correspond seulement à un rapport de 1,36. Pourtant, le changement réel de température est identique dans les deux cas. Le tableau 1.2 montre aussi clairement le caractère arbitraire du point zéro qui ne correspond pas aux mêmes températures sur les échelles Celsius et Fahrenheit. En définitive, calculer des rapports entre scores n'a de signification que pour les échelles de rapports. Ainsi par exemple, on sait qu'une personne qui pèse 100 Kg a exactement deux fois le poids d'une personne de 50 Kg.

Encart 1.1 Les échelles en psychologie

En psychologie et même en neurosciences, les études scientifiques impliquent souvent de demander aux sujets de répondre à des questionnaires permettant par exemple d'évaluer leurs attitudes, de mesurer leurs émotions et leurs comportements ou de caractériser leurs traits de personnalité. Parmi les différents types de questions posées, on retrouve très fréquemment des échelles de mesure dont l'objectif est de quantifier le degré d'accord avec une affirmation particulière. Il existe différents types d'échelle de mesure dont la description détaillée dépasse largement le cadre du présent ouvrage. Cependant, force est de constater que les échelles de Likert sont les plus connues et les plus fréquemment utilisées.

Une échelle de type Likert, ainsi dénommée d'après le psychologue américain Renis Likert (1903-1981) qui en a popularisé l'usage, est une échelle graduée comprenant cinq valeurs comprise entre « complètement d'accord » et « pas du tout d'accord » et permettant de quantifier le degré d'accord envers une affirmation. Toutefois, dans la pratique, le terme « échelle de Likert » est un peu abusivement étendu à toutes les échelles graduées sur lesquelles il n'est pas possible de choisir des valeurs non-entières. Les échelles de ce type sont ainsi souvent graduées en sept points comme dans l'exemple suivant :

Je peux me faire facilement des amis :

1 2 3 4 5 6 7

Pas du tout d'accord Tout-à-fait d'accord

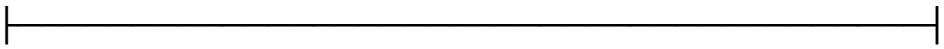
Le répondant doit entourer l'un des chiffres (d'où l'impossibilité de valeurs non-entières) pour marquer son degré d'accord envers l'affirmation « Je peux me faire facilement des amis ».

La manière de traiter les données obtenues sur des échelles de type Likert a fait l'objet de nombreux débats, en particulier en ce qui concerne leur niveau de mesure, ordinal ou métrique. Le problème principal que posent ces échelles est la constance des unités. Rien n'indique en effet que l'écart entre 2 et 3 représente nécessairement la même différence que l'écart entre 5

et 6 par exemple. Bien que certains auteurs continuent à traiter les échelles de Likert comme des variables métriques, il semble dès lors plus prudent de les traiter préférentiellement comme des variables ordinales.

Afin d'améliorer la précision des mesures et de considérer plus facilement les réponses comme une variable métrique (bien que cette dernière affirmation puisse être discutée), on recommande généralement de remplacer les échelles de type Likert par des échelles visuelles analogiques. L'échelle visuelle analogique se présente le plus souvent sous la forme d'un segment de droite horizontal de 10 cm dont les extrémités sont définies comme les limites du paramètre à mesurer. La tâche du répondant consiste à placer une croix sur la ligne à l'endroit qui définit le mieux sa réponse entre les deux opposés. L'exemple précédant transformé en échelle visuelle analogique donnerait :

Je peux me faire facilement des amis :



Pas du tout d'accord Tout-à-fait d'accord

La réponse du sujet est alors encodée en mesurant précisément la distance en millimètres à partir du point zéro (ici pas du tout d'accord).

L'avantage des échelles visuelles analogiques par rapport aux échelles de Likert est de permettre des réponses plus précises qui peuvent être considérées comme métriques. Elles sont toutefois plus laborieuses à encoder puisque l'analyse des résultats d'un questionnaire papier nécessite l'utilisation d'une latte par le chercheur. Imaginez le travail pour analyser un questionnaire contenant 25 échelles analogiques et administré à 200 sujets. Heureusement, grâce au développement des questionnaires en ligne, l'encodage des réponses aux échelles visuelles analogiques peut être réalisé automatiquement par un ordinateur.

Hiérarchie dans le niveau de mesure des variables

Entre les trois types de variables définies ci-dessus (nominale, ordinale et métrique), il y a plus qu'une distinction, il y a une réelle hiérarchie dans le niveau de mesure et de précision (voir figure 1.1). Les trois catégories de variables peuvent être classées sur un continuum en fonction de la quantité d'information fournie. Une variable nominale fournit moins d'information qu'une variable ordinale, qui donne à son tour moins d'information qu'une variable métrique. Les trois types de variable permettent de dire si deux valeurs sont semblables ou différentes. Pour les variables nominales, c'est d'ailleurs la seule information fournie. Les variables ordinales permettent en outre de dire si une valeur est égale, supérieure ou inférieure à une autre. Elle introduit donc une information supplémentaire de hiérarchie entre les valeurs. On dira donc que les variables ordinales ont un niveau de mesure supérieur aux variables nominales. Enfin, les variables métriques proposent des unités standards qui fournissent une information précise (en unités) sur la différence entre deux scores. Ce sont les variables métriques qui fournissent le plus d'information et qui ont donc le niveau de mesure le plus élevé.

La hiérarchie dans le niveau de mesure correspond aussi au degré de sophistication des traitements statistiques qui pourront être effectués avec ces variables. Des variables métriques pourront ainsi être traitées avec des analyses plus sophistiquées que les variables ordinales et nominales. Ce sont les variables nominales qui autorisent les traitements les moins sophistiqués. Pour anticiper sur une question que nous aborderons dans un chapitre ultérieur, on peut également dire que la puissance statistique évolue parallèlement au niveau de mesure des variables. Les traitements statistiques réalisés sur des variables nominales sont moins puissants que ceux effectués sur les variables ordinales, qui sont à leur tour moins puissants que ceux que l'on peut accomplir avec des variables métriques. C'est dire s'il est important de choisir correctement le niveau de mesure des variables que l'on souhaite étudier. Lors de la planification d'une étude, il est toujours préférable de définir une variable en utilisant le niveau de mesure le plus élevé possible. A l'issue de l'étude, lors du traitement des données, il sera toujours possible de recoder une variable en utilisant un niveau de mesure moins élevé. Par contre, l'inverse n'est pas vrai. Il est en effet impossible de recoder une variable d'un niveau de mesure bas vers un niveau plus élevé.

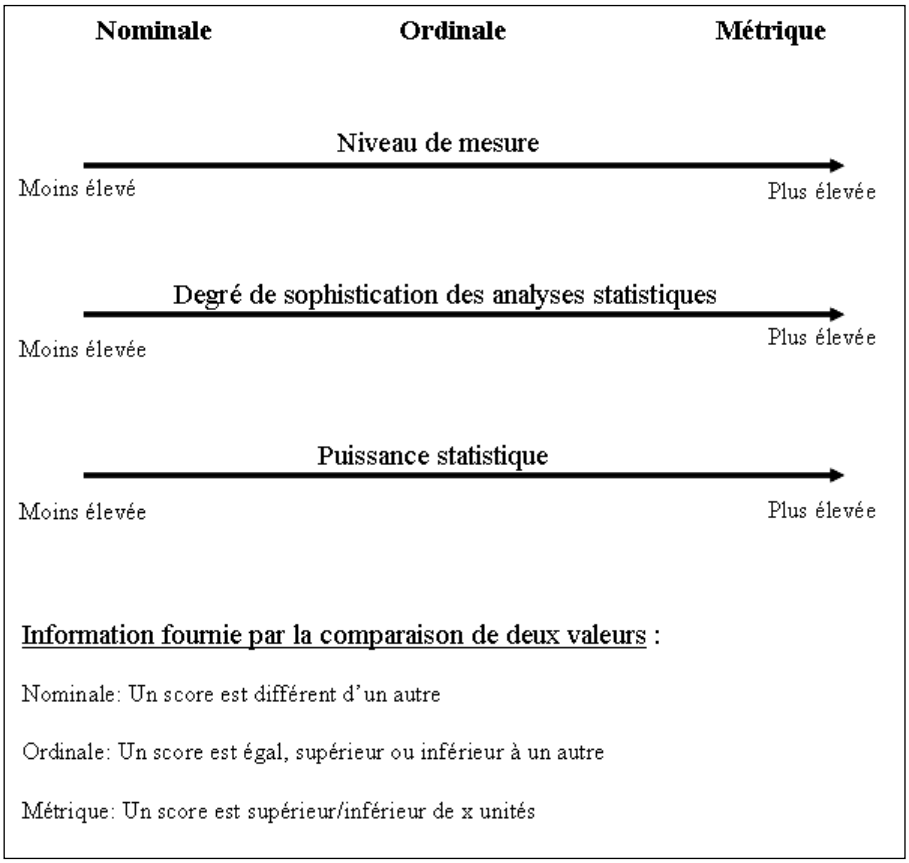


Figure 1.1 – Hiérarchie dans le niveau de mesure des variables

Considérons l'exemple d'une étude dans laquelle on souhaite enregistrer l'âge des participants. Voici trois manières de les interroger sur leur âge qui correspondent aux trois niveaux de mesure :

Tableau 1.3 – Exemples de questions pour mesurer l'âge des sujets dans une étude

Mesure nominale	Mesure ordinale	Mesure métrique
Vous êtes (cochez une case) : □ Mineur (moins de 18 ans) □ Majeur (plus de 18 ans)	Vous avez (cochez une case) : □ Moins de 20 ans □ Entre 20 et 40 ans □ Entre 40 et 60 ans □ Plus de 60 ans	Votre âge est de…….. ans

L'âge enregistré avec la variable métrique fournit plus d'information et pourra si nécessaire être recodée ultérieurement selon les quatre catégories de la variable ordinale ou selon les deux modalités de la variable nominale. Par contre, si l'âge a été enregistré sous forme de variable ordinale, il ne sera plus possible de retrouver l'âge exact des participants après la fin de l'étude et donc de recoder les données en variable métrique. Dans ce cas, les traitements statistiques qui pourront être effectués sur la variable « âge » seront limités au degré de sophistication qu'autorisent les variables ordinales. Toutes les variables n'offrent évidemment pas le choix entre les trois niveaux de mesure. Par exemple, s'il s'agit de noter la profession des participants, seule une variable nominale le permettra.

Pour aller plus loin : Variables discrètes et continues

A côté de la distinction entre les variables nominales, ordinales et métriques, les variables sont aussi traditionnellement classées en variables discrètes et continues. Cependant, cette distinction est moins déterminante pour le choix des traitements statistiques appropriés aux données et peut donc être ignorée. Toutefois, comme la majorité des traités de statistique font cette distinction, nous allons la présenter brièvement. Une ***variable continue*** *peut prendre n'importe quelle valeur à l'intérieur d'un intervalle de valeurs possibles, y compris des fractions de l'unité standard de mesure. Prenons un exemple classique de la psychologie cognitive pour illustrer le propos. En psychologie cognitive, beaucoup d'études sont basées sur la mesure du temps de réaction à divers stimuli. A condition de disposer d'un instrument de mesure suffisamment précis, n'importe quel score de temps de réaction sera possible, y compris des fractions de millisecondes. On pourrait par exemple mesurer un temps de réaction de 254,68 ms ou de 389,12 ms. Une* ***variable discrète****, par contre, ne peut prendre que certaines valeurs précises à l'intérieur de l'ensemble des valeurs possibles. Les variables discrètes ont donc un nombre limité de valeurs possibles qui peuvent être énumérées explicitement. Pour faire le lien avec les niveaux de mesure décrits ci-dessus, une variable nominale est toujours discrète, une variable ordinale est le plus souvent discrète, tandis qu'une variable métrique peut être soit discrète soit continue. Le nombre de fois qu'une personne a consommé du cannabis lors des six derniers mois est un exemple de variable métrique discrète. Une personne pourra répondre 4 fois, 50 fois ou 100 fois, mais pas 4,34 fois (les fractions d'unité sont impossibles). Pour les variables métriques, la distinction entre variables discrètes et continues est parfois malaisée. En effet, les instruments de mesure sont souvent imprécis et ne permettent que des valeurs arrondies à l'unité. On croit alors travailler avec une variable discrète alors*

que la variable sous-jacente est en réalité continue. Prenons l'exemple de la variable « poids d'un sujet ». Même si la balance utilisée ne permet qu'une mesure arrondie à l'unité (kilogramme), il s'agit d'une variable continue. Avec une balance plus précise on pourrait peser une personne avec une précision au centième voire au millième de kilogramme. Théoriquement, toutes les valeurs de poids sont possibles, même si l'instrument de mesure ne permet pas une telle précision.

1.3 Variables indépendantes et dépendantes

Un des objectifs majeurs de la statistique consiste à mettre en évidence l'existence de relations entre variables. Plus largement, la majorité des études expérimentales visent à mettre en évidence des relations de causalité entre variables. A cet effet, on distingue généralement les variables indépendantes et les variables dépendantes.

Dans une étude expérimentale, on fait le plus souvent varier une chose et on observe l'effet produit sur une autre. La variable manipulée par l'expérimentateur est la variable indépendante. La variable dépendante est celle qui change en fonction de la variable indépendante. Les différentes valeurs de la variable dépendante sont les données (les résultats) de l'expérience ou de l'étude. On ne peut pas agir directement sur ces valeurs, elles sont simplement collectées.

Conceptuellement, la variable indépendante est celle à laquelle on attribue une nature causale dans l'étude et la variable dépendante, celle qui constitue l'effet ou la conséquence, c'est-à-dire qui varie en fonction de la cause.

Il est important de bien distinguer variables indépendantes et variables dépendantes car beaucoup de procédures statistiques traitent différemment les deux types de variables. En règle générale, le type des variables (nominale, ordinale ou métrique) identifiées comme dépendante et indépendante dans une étude particulière déterminera la technique statistique qui sera utilisée pour traiter les données. Avec certaines procédures statistiques (voir par exemple la régression linéaire, chapitre 3), des résultats différents seront obtenus si on intervertit les variables considérées comme indépendante et dépendante dans l'analyse. Malheureusement, dans certaines études, il est parfois bien difficile d'identifier une variable qui devrait être considérée comme indépendante et une autre qui serait dépendante. Par exemple, si on étudie la relation entre anxiété et dépression, il est difficile de déterminer laquelle des deux variables doit être

considérée de nature causale. Heureusement, certaines procédures statistiques (par exemple, la corrélation, chapitre 3) ne font pas cette distinction entre variable indépendante et variable dépendante.

1.4 Erreurs typiques lors des traitements statistiques

Lors de l'analyse des données d'une étude, les erreurs les plus fréquentes et les plus graves portent souvent sur la matière de ce premier chapitre. Il s'agit bien souvent d'une mauvaise identification des unités d'observation ou d'une mauvaise définition des types de variable en jeu. Si les unités d'observation ou les variables sont mal définies, les conséquences seront généralement désastreuses. Cela conduira souvent à choisir des traitements statistiques inadéquats qui ne permettent pas de répondre aux questions et hypothèses posées et parfois même à réaliser des analyses qui n'ont aucun sens. Voici quelques exemples d'erreurs fréquemment rencontrées :

1. Mauvaise définition ou identification de l'unité d'observation

Une erreur classique dans l'analyse des données d'une étude consiste à identifier incorrectement les unités d'observation. **Les unités d'observation sur lesquelles porte l'étude sont essentiellement indépendantes les unes des autres**. En d'autres termes, les résultats obtenus sur une variable par une unité d'observation ne doivent avoir aucun lien avec les résultats d'une autre unité d'observation. Prenons deux exemples pour illustrer un type d'erreur classique.

Dans une étude de psychologie, plusieurs groupes de quatre sujets sont formés. On demande à chacun des groupes de débattre d'une question hautement polémique et on enregistre le nombre d'attaques verbales émises par chacun des sujets. L'hypothèse testée est que le nombre d'attaques verbales sera moindre dans les groupes mixtes comprenant des garçons et des filles que dans les groupes unisexes. Dans cette étude, la véritable unité d'observation n'est pas le sujet mais le groupe. A l'intérieur d'un groupe, le nombre d'attaques verbales émises par un sujet n'est pas indépendant de celles des autres sujets du même groupe. Il est évident qu'il y a une dynamique à l'intérieur de chaque groupe qui détermine le niveau plus ou moins élevé des attaques verbales. Dans un cas comme celui-ci, l'unité d'observation principale est le groupe, même si certains traitements statistiques sophistiqués permettent de traiter les résultats de chaque sujet comme des sous-unités à l'intérieur des groupes.

En neurobiologie, on rencontre aussi très souvent des problèmes d'identification des unités d'observation. Par exemple, il est fréquent de prélever plusieurs échantillons sur chacun des sujets de l'étude. Ce sera le cas si on décide de prélever plusieurs échantillons sur des cerveaux de rats qui auraient été soumis à un traitement expérimental particulier. Il serait alors erroné de considérer chaque échantillon comme une unité d'observation indépendante. C'est en réalité le rat qui constitue l'unité d'observation de base. Les différents échantillons prélevés sur un même sujet devront être traités comme des mesures répétées sur la même unité d'observation.

2. Mauvaise identification du niveau de mesure d'une variable

Il est assez fréquent, particulièrement chez les étudiants, d'observer des confusions entre variables nominales, ordinales et métriques. Dans certains cas, de telles confusions n'auront pas de graves conséquences. Par exemple, une variable ordinale chiffrée pourra souvent être traitées comme le serait une variable métrique. Dans d'autres cas, les conséquences pourront être dramatiques. Par exemple, appliquer des traitements statistiques pour variables métriques à une variable nominale conduira immanquablement à des conclusions absurdes. Ce genre d'erreur est d'autant plus fréquent que beaucoup de logiciels statistiques permettent d'un simple clic de réaliser n'importe quelle analyse sur n'importe quelle variable. Afin d'éviter ce type d'erreur, il est souvent utile de se demander quelle information est fournie par la comparaison des scores de deux sujets sur une variable (voir figure 1.1). L'identification de la variable est alors effectuée sur base de la réponse à cette question : deux scores sont seulement semblables ou différents (variable nominale), deux scores peuvent être classés selon un axe supérieur-inférieur (variable ordinale), la différence entre deux scores peut être précisément définies en termes d'unités standards (variable métrique).

3. Inversion des variables indépendantes et dépendantes

Certains traitements statistiques nécessitent de définir précisément la ou les variables indépendantes et dépendantes d'une étude. Pour ce type d'analyse, une confusion entre les variables indépendantes et dépendantes conduit le plus souvent à des erreurs d'interprétation dont la gravité est variable selon le contexte de l'étude. Rappelons toutefois que certaines analyses statistiques ne définissent pas de variables indépendantes et dépendantes (par exemple la corrélation).

En résumé, les traitements statistiques reposent sur le concept de variable. Une variable définit une propriété d'une unité d'observation (d'un sujet) qui peut prendre différentes valeurs. Chaque variable assigne une et une seule valeur à chacune des unités d'observations de l'étude. Choisir correctement le traitement statistique à appliquer aux données d'une étude implique :

1. D'identifier correctement les unités d'observations indépendantes de l'étude. Il s'agit souvent de sujets mais pas toujours.
2. De définir le niveau de mesure (nominal, ordinal ou métrique) des variables.
3. Dans certains cas, de définir quelles sont les variables indépendantes et dépendantes de l'étude.