

3.1.4 - Organiser les données

3.1.4.1 Définir la nature des observations

3.1.4.2 Caractériser le jeu de données

3.1.4.3 Construire la matrice des données

3.1.4.4 Types de tableaux obtenus

Tout traitement informatique ou manuel des données nécessite au préalable de connaître le type de jeu de données que l'on utilise. La première chose à faire est de parcourir les données pour juger de l'importance du travail à effectuer, et découvrir quels types de variables entrent en jeu, quels traitements seront à effectuer. Le poids d'un individu, la couleur des pétales d'une fleur ne sont pas des données de même nature. Il est important de reconnaître les types de variables car «à chaque variable correspond ses calculs» (voir chapitre 3.2.).

3.1.4.1 Définir la nature des observations

Il est important de considérer la nature des données (observations) que l'on va traiter. Les observations peuvent être soit quantitatives, soit qualitatives.

Variables quantitatives

On désigne par le vocable variable quantitative toute série de chiffres se rapportant à une quantité.

Les données quantitatives comprennent les dénombrements (ou comptages) et les mesures (ou mensurations). Dans le cas des dénombrements, la caractéristique étudiée est une variable discrète ou discontinue, ne pouvant prendre que des valeurs entières non négatives (nombre de pétales par fleur...). Il suffit de compter le nombre d'individus affectés par chacune des valeurs (fréquences) de la variable.

Dans le cas de mesures, la variable est de nature continue (hauteur, poids, surface, concentration, température...). Les valeurs possibles sont illimitées mais du fait des méthodes de mesure et du degré de précision de l'appareil de mesure, les données varient toujours de façon discontinue.

UNE VARIABLE

D'après SCHERRER (1984) une variable est une caractéristique mesurée ou observée sur chacun des éléments de l'échantillon ou sur des entités prédéfinies qui se rattachent aux unités d'échantillonnage.

En effet, il peut s'agir d'un attribut de l'élément comme la taille, le poids d'un animal, on parlera alors de variables propres. S'il s'agit d'une composante particulière de son environnement comme la quantité de nourriture disponible, la température ambiante, etc., on parlera alors de variable associée, car elle n'est pas mesurée sur l'élément proprement dit.

Les variables peuvent être de nature qualitative ou quantitative.

Une variable, est de nature qualitative si elle ne peut être mesurée tout en demeurant susceptible de classement, comme la classe d'âge (adulte ou immature), le sexe, la race, l'espèce...

Une variable est de nature quantitative si elle peut être mesurée, comme la hauteur d'un arbre, le poids d'un micromammifère, son nombre de parasites,... (voir aussi chapitre 3.1.4.1. : «Définir la nature des observations»).

Variables qualitatives

On désigne par le vocable variable qualitative toute série de chiffres se rapportant à des qualités (voir aussi encadré). Les données qualitatives peuvent être assimilées au cas des variables discontinues, en supposant que les différentes variantes du caractère qualitatif sont rangées dans un ordre correspondant par exemple à la suite des nombres entiers positifs (différentes couleurs, différents degrés d'infection...).

3.1.4.2 Caractériser le jeu de données

Ensuite, il faudra chercher à caractériser l'ensemble des données car les traitements statistiques

différent d'un jeu de données à l'autre.

Série statistique simple

Une série statistique simple est un ensemble de données relatif à une variable mesurée sur un échantillon ou une population d'éléments (SCHERRER, 1984). Exemple : la taille des individus d'une espèce.

Série statistique double

Une série statistique double est un ensemble de couples de données relatif à la même mesure de deux variables sur un échantillon ou une population d'éléments. Comme ces deux variables peuvent être indifféremment qualitatives ou quantitatives, il est possible de distinguer trois cas de figure :

- les deux variables sont quantitatives. Il s'agit, par exemple, d'une série d'individus dont on mesure la taille et le poids,
- les deux variables sont qualitatives : par exemple le sexe et le comportement d'une série d'individus,
- une variable est qualitative et l'autre quantitative : poids et sexe d'une série d'individus. Par exemple lors d'un suivi d'oiseaux par bagueage, on relève sa taille, son poids, la longueur ailaire, mais aussi sa provenance (oiseau bagué)...

Série statistique multiple

Le nom de série statistique multiple est réservé aux séries comportant plus de deux variables. Il s'agit donc d'un ensemble de données relatives à la mesure de «p» variables sur un échantillon de «n» éléments ou une population de «N» unités.



Photo n°52 : Mesure de la taille du bec d'un Fou de Bassan dans la RN des Sept-Iles. © LPO ÎLE GRANDE.

3.1.4.3 Construire la matrice des données

Lors de la collecte des données, les valeurs observées se trouvent évidemment sans ordre. Si le volume d'informations devient encombrant, il devient très vite difficile de dégager les caractéristiques ou les faits importants de cette accumulation de chiffres. L'ensemble des résultats bruts acquis et saisis dans la base de données doit être structuré pour faire avancer la compréhension d'un phénomène et réaliser des graphiques (voir chapitre 3.2.1) et des tests statistiques (voir chapitre 3.2.3).

Organiser les données

L'organisation des données dans un tableau sous un format particulier est nécessaire pour le calcul de divers indices ou encore la détermination des associations végétales par le regroupement et la comparaison des données issues de plusieurs relevés (tableaux phytosociologiques de Braun-Blanquet par exemple).

Le gestionnaire procèdera au regroupement ou au classement des observations, dans un tableau à double entrée appelé «matrice des données» (tableau de distribution de fréquence par exemple) à partir duquel il devient possible de les représenter graphiquement. Dans ce tableau, chaque ligne de résultats correspond aux caractéristiques d'un élément, et chaque colonne aux différentes variantes de chacune des variables (voir Figure n°56).

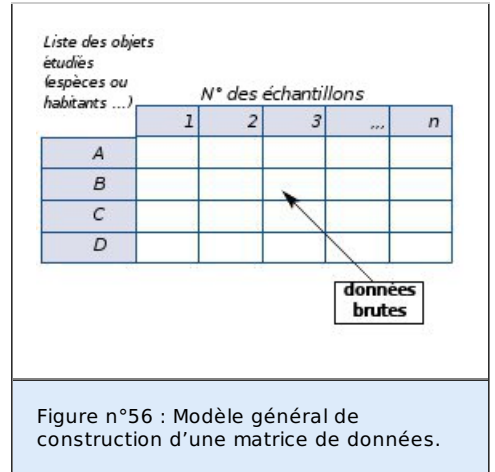


Figure n°56 : Modèle général de construction d'une matrice de données.

Recommandations

LA FRÉQUENCE ABSOLUE D'UNE CLASSE
C'est simplement l'effectif où le nombre d'éléments appartenant à cette classe.

Ces tableaux peuvent être réalisés à l'aide du logiciel Excel, qui permet des analyses croisées de jeux de données (notamment après exportation des données à partir de la base de données servant à la saisie, notamment à l'aide du logiciel SERENA de RNF (voir page 73).

LA FRÉQUENCE RELATIVE D'UNE CLASSE
C'est le rapport de son effectif à l'effectif total de l'échantillon.

Dans la plupart des cas, il est nécessaire de procéder à un traitement préalable adéquat du tableau initial (codages particuliers...).

Si la fréquence d'observation n'a pas été respectée, il est possible de calculer les données des échantillons manquants (voir encadré ci-contre).

Si l'effectif de l'échantillon est faible, il n'y a pas d'inconvénient à conserver les données les unes à la suite des autres sans les classer.

3.1.4.4 Types de tableaux obtenus

Pour effectuer un tel classement, on distingue les séries statistiques simples des séries statistiques doubles, (voir chapitre 3.1.4.2) (SCHERRER, 1984).

Le cas des séries statistiques simples Dans le cas de séries statistiques simples, on réalise un tableau de distribution de fréquence. C'est un mode synthétique de présentation des données numériques, montrant comment les résultats enregistrés sur une variable se distribuent dans les différentes classes d'une échelle de valeurs ou dans différentes catégories qualitatives (SCHERRER, 1984). Pour préparer un tableau de distribution de fréquence, il faut déterminer des classes et compter le nombre d'éléments appartenant à chacune d'elles (voir Tableau n°16 et n°17). La construction des tableaux de distribution de fréquence nécessite donc le calcul de fréquences.

C'est ce type de tableau qui est obtenu à partir de la plupart des données recueillies dans les réserves naturelles.

Tableau n°16 : Répartition du Grand tétras en fonction de l'exposition dans le Massif du Madres Coronat (RN de Nohèdes).

Exposition	Nombre de données (%)	Nombre d'indices observés (%)
ns	14,89	10,97
w	2,13	0,65
nw	1,06	0,65
n	28,72	35,16
ne	22,34	27,42
e	15,96	18,39
ese	1,06	0,32
se	6,38	3,55
s	7,45	2,9

Source : MOREAU, 1996

Tableau n°17 : Proportion des différents types d'indices récoltés, sur l'année et par saison.

Type d'indice	abondance relative (%)				
	Année	Hiver	Printemps	Été	Automne
animal vu	9,54%	7,63%	10,52%	16,66%	6,35%
croûte	44,68%	42,13%	53,11%	43,25%	41,68%
poils	0,06%	0,11%	0,05%	0,00%	0,00%
trace ou piste sur neige	20,34%	32,88%	7,72%	0,46%	26,12%
trace ou piste sur boue	4,73%	3,37%	6,51%	6,91%	3,66%
cadavres/os	0,17%	0,14%	0,20%	0,20%	0,16%
dégâts aux végétaux	4,90%	2,72%	4,86%	6,32%	7,97%
relief de repas	6,03%	5,30%	6,86%	7,24%	5,55%
activité sur sol	7,71%	5,08%	8,42%	14,02%	6,84%
terrier/habitat	1,40%	0,53%	1,15%	3,49%	1,62%
autre type d'indice	0,44%	0,11%	0,60%	1,45%	0,05%

Source : LETSCHER, 2001

Dans ce type de tableau, on présentera éventuellement les variables par ordre décroissant. On pourra ainsi faire ressortir un effet de taille (voir Tableau n°24).

Le cas des séries statistiques doubles

Dans le cas des séries statistiques doubles avec deux variables quantitatives, on réalise un tableau dit de corrélation. Connaissant les bornes de chaque classe sur chacune des variables, on construit un tableau à double entrée portant les indices de classe de la variable x dans le sens horizontal et de la variable y dans le sens vertical. Pour chacune des cellules correspondant aux intersections des lignes et des colonnes, on dénombre les éléments appartenant à cette classe (SCHERRER, 1984).

Si on étudie simultanément deux caractéristiques qualitatives, les résultats sont classés en un tableau de contingence. Dans ce tableau à double entrée, les colonnes correspondent aux différentes catégories de classement de la première variable qualitative, et les lignes aux catégories de classement de la deuxième variable. Dans chacune des cellules coïncidant avec les intersections des colonnes et des lignes, on porte le nombre d'éléments appartenant aux catégories correspondantes (voir Tableau n°18).



Photo n°53 : Accenteur mouchet à la RN de l'Étang de la Mazière.
© A. DAL MOLIN/L. JOUBERT.

Tableau n°18 : Tableau de contingence. Bilan de prélèvements de sangliers dans la RN des Manneville (février 2002)

Age	Mâles	Femelles	Total
M arcassin	0	0	0
6 mois - 1 an	11	7	18
1 an - 2 ans	1	2	3
2 ans - 3 ans	1	1	2
3 ans - 4 ans	0	3	3
4 ans et plus	0	1	1
Total	13	14	27

Source : Bilan 2002 de la réserve naturelle des Manneville.

Les données de séries statistiques doubles avec une variable qualitative, l'autre quantitative, sont elles aussi classées dans un tableau à double entrée, qui résulte de la combinaison d'un tableau de contingence et d'un tableau de corrélation (voir Tableaux n°19 et n°20).

Tableau n°19 : Tableau de synthèse des masses enregistrées chez l'Accenteur mouchet dans la RN de l'Étang de la Mazière.

Masses enregistrées	Individus de 1A		Individus adultes	
15 grammes	2	0,35%	0	0,00%
16 grammes	14	2,43%	2	1,38%
17 grammes	81	14,08%	12	8,33%
18 grammes	148	25,74%	35	24,31%
19 grammes	141	24,52%	49	34,03%
20 grammes	108	18,78%	22	15,27%
21 grammes	45	7,82%	15	10,42%
22 grammes	16	2,78%	7	4,86%
23 grammes	15	2,61%	0	0%
24 grammes	3	0,52%	2	1,38%
25 grammes	0	0%	0	0%
26 grammes	2	0,35%	0	0%
	575		144	

Source : DAL MOLIN et al., non daté

Tableau n°20 : Distribution de fréquence des longueurs d'ailes de catégories différentes de sexe du Pipit spioncelle dans la RN de l'Étang de la Mazière.

	81 mm	95 mm	96 mm	97 mm	98 mm	TOTAL
Femelles présumées	2 soit 0,21 %					2
Mâles présumés		35 soit 3,78 %	11 soit 1,19 %	5 soit 0,54 %	3 soit 0,32 %	54
						56 soit 6,06 %

Source : DAL MOLIN et al., non daté

Le cas des séries statistiques multiples

Les données issues de séries statistiques multiples qui feront l'objet de représentations graphiques (voir pages 80 et 81) sont elles aussi présentées sous forme matricielle, c'est-à-dire d'un tableau composé de «p» colonnes et de «n» lignes (SCHERRER, 1984) pour permettre une analyse factorielle (voir page 81). Le tableau n°21 est par exemple réalisé à partir du tableau de données brutes n°15.

Tableau n°21 : Matrice de corrélation entre les variables étudiées dans l'ACPI à la RN des Marais de Bruges.

	MVS	HT	V	R	OMB	T°	PH	O2MG/L	O2%	C	IP	NO3	NO2	NH4	PCN	P	NK	CHL	PHEO	MES
MVS	1																			
HT	-0,16	1																		
V	-0,06	-0,3	1																	
R	0,04	-0,22	0,15	1																
OMB	0,13	0,26	-0,1	-0,28	1															
T°	-0,09	-0,13	0,07	0,52	-0,01	1														
PH	-0,02	-0,09	-0,18	0,05	0,29	0,43	1													
O2MG/L	-0,09	-0,24	-0,25	-0,38	-0,15	-0,04	0,47	1												
O2%	-0,05	-0,29	-0,21	0	-0,15	0,37	0,59	0,89	1											
C	-0,03	-0,19	-0,03	-0,54	-0,06	-0,56	-0,27	0,30	0,03	1										
IP	0,19	-0,34	-0,52	-0,25	-0,22	-0,27	0,31	0,4	0,17	0,20	1									
NO3	-0,04	0,16	-0,07	-0,16	-0,01	0,05	-0,03	-0,16	-0,23	-0,06	0,09	1								
NO2	0,04	0,19	-0,01	-0,18	0,3	0,38	0,34	-0,03	0,06	-0,12	-0,13	0,54	1							
NH4	0,21	0,09	0,02	-0,34	0,12	-0,38	-0,23	0,16	-0,05	0,21	0,09	0,03	-0,17	1						
PCN	-0,27	0,11	0,01	-0,02	0,1	0,12	0	-0,35	-0,37	-0,12	-0,26	0,03	0,46	-0,14	1					
P	-0,2	0,25	0,23	-0,04	0,3	0,14	-0,28	-0,22	-0,22	-0,11	-0,32	0,23	0,53	-0,15	0,67	1				
NK	0,39	-0,11	-0,04	-0,21	-0,22	-0,26	0	0,15	-0,05	0,20	0,55	0,04	-0,34	0,53	-0,19	-0,27	1			
CHL	0,54	0,05	-0,12	-0,16	0,27	0,02	0,05	0,03	0,08	0,07	0,05	-0,31	0,21	0,27	-0,31	-0,15	0,21	1		
PHEO	0,61	-0,05	0,07	0,03	0,15	0,07	0	-0,18	-0,07	-0,16	-0,01	-0,29	-0,05	0,10	-0,37	-0,14	0,24	0,58	1	
MES	0,59	-0,15	0,01	0,04	0,07	-0,13	0,01	-0,02	-0,05	0,01	0,33	-0,04	0,06	0,28	-0,23	-0,14	0,49	0,59	0,0	1

Source : MERIGON et al., 1999

[Haut de page](#)

Tous droits réservés © - Propriété de l'AFB