Organiser et représenter des données

J-M., GALHARRET

LMJL, Faculté de Psychologie, Nantes Université

Introduction

Introduction

- En psychologie, on doit mesurer sur des individus de nombreuses caractéristiques (appelées variables en statistique).
- La finalité est souvent d'établir des liens entre ces variables (ex : existe-t-il un lien entre le fait d'être dépressif et l'âge ?)
- Cette année on va commencer par identifier la nature des variables, par les représenter graphiquement, par les décrire et les résumer.

Base de données

Une base de données se présentera toujours sous la forme d'un tableau :

- Les colonnes du tableau représentent les caractéristiques (variables) des participants.
- Les lignes du tableau représentent les valeurs correspondantes à chacun des participants.

Exemple : Données issues d'une étude durant le premier confinement (2020)

participant_ID	Sexe	diag_COVID_C1	Vuln1	Conf_OMS_C1	Act_PRO
ll00ta87as	1	0	13	3	2
LE90AN90EN	2	0	10	4	1
Le00ny93lt	2	0	50	3	3
de12se96ot	2	0	88	3	1
du30an01le	2	0	2	4	3
mo00va85ne	1	1	100	3	1
BA20NE66UE	1	1	100	3	1
DE00NE85TE	1	1	100	3	1

Remarque :

Dans l'étude on avait 2389 participants et 145 caractéristiques.

Type de variables, échelles

Variables catégorielles et numériques

On distingue deux grandes familles de variables :

- Les variables **catégorielles ou qualitatives**. Les participants sont placés dans des catégories indépendantes et séparées (Ex : CSP,).
- Les variables **numériques ou quantitatives**. Les participants sont placés sur une échelle qui possède une direction.

Les **modalités** d'une variables sont toutes les valeurs que peut prendre la variable.

ATTENTION : Les modalités d'une variable catégorielle peuvent être codées numériquement, elles ne sont pas pour autant des variables numériques !

Les échelles de mesure :

Type variable	Type données	Exemple
Catégorielle ou qualitative	Nominale	Sexe, CSP
Catégorielle ou qualitative	Ordinale	Niv d'étude, Niv d'accord
Numérique ou quantitative	Discrète	Nb d'enfants, Nb d'erreurs
Numérique ou quantitative	Continue	Temps de réponse, Taille, Score

Remarque : Il arrive souvent qu'on oublie le type de variable et on va dire **variable nominale** au lieu de **variable catégorielle mesurée sur une échelle nominale**.,

Exemple

Pour chacune des questions suivantes, quelle est la nature de la variable associée ?

Q44. Avez-vous été testé·e/reconnu·e positif au COVID-19 ?

- ○ Oui
- ◉ Non
- ○ Ne souhaite pas répondre

Confiance.
Quel est votre degré de confiance envers les institutions et personnalités suivantes pour définir les actions à mener face à la crise du COVID-19 ?

	1 Pas du tout confiance	2	3	4	5 Tout à fait confiance
L'Organisation Mondiale de la santé	○	◉	○	○	○

Réponse

- Question 1 : on lui associe une variable nommée diagnost_COVID_C1 qui prendra les modalités OUI/ NON. Il s'agit d'une variable catégorielle à 2 modalités (on dit dans ce cas que la variable est **binaire** ou **dichotomique**).
- Question 2 : on lui associe une variable nommée Vuln1 qui prendra n'importe quelle valeur entre 0 et 100. Il s'agit d'une variable numérique sur une échelle **continue**.
- Question 3 : variable nommée CONF_OMS_C1 qui est catégorisée sur une échelle **ordinale** en 5 points.

Variables ordinales et variables discrètes :

- Une variable numérique peut aussi être mesurée sur une échelle discrète (ex : nombre d'enfants, nombre d'erreurs, nombre de pièces dans l'habitation...).
- La description des variables numériques sur échelle discrète est proche de celle d'une variable catégorielle sur échelle ordinale. On peut cependant utiliser des indices résumés avec une variable discrète alors que l'on ne peut pas le faire avec une variable oridinale.
- La variable discrète comporte souvent plus de modalités que la variable ordinale.

Décrire et représenter une variable catégorielle

Décrire une variable catégorielle

- On compte le nombre de participants n_i correspondants à chaque modalité i (n_i : effectifs). On note N le nombre total de participants
- On calcule la proportion de participants f_i (appelée **fréquence**) correspondants à chaque modalité. On a

$$f_i = \frac{n_i}{N},$$

la fréquence est souvent exprimée sous forme de % (on multiplie le résultat précédent par 100)

Exemple 1 : diagnost_COVID_C1

modalité	effectif	fréquence (%)
NON	2356	98.62
OUI	24	1.00
NA	9	0.38

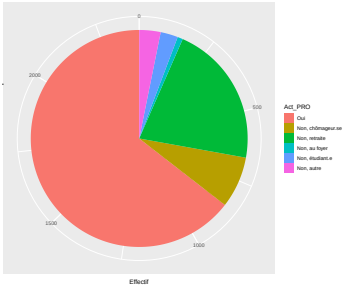
Exemple 2 : activité professionnelle des participants

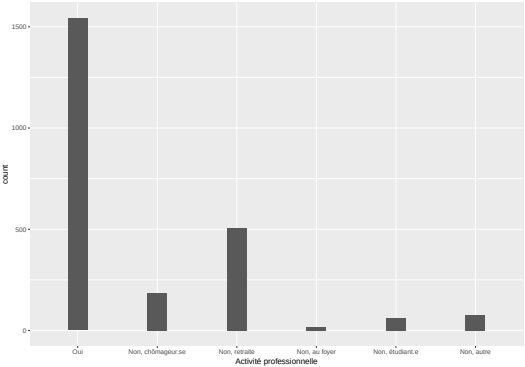
modalité	effectif	fréquence (%)
Oui	1541	64.50
Non, chômageur.se	183	7.66
Non, retraite	508	21.26
Non, au foyer	19	0.80
Non, étudiant.e	62	2.60
Non, autre	76	3.18

Représenter une variable catégorielle

- Diagramme circulaire (camembert) : l'aire des secteurs angulaires est proportionnelle à l'effectif de chaque modalité.
- Diagramme en bâtons : la hauteur des bâtons est proportionnelle à l'effectif de chaque modalité.

Exemple avec activité professionnelle





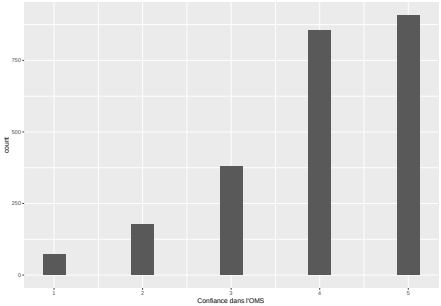
Représenter une variable ordinale :

Pour les variables ordinales on peut également utiliser des diagrammes en bâtons (même principe que les variables nominales).

Exemple avec Confiance dans l'OMS

CONF_OMS_C1_2	Effectif	Fréquence
1- Pas du tout confiance	72	3.0
2- Peu confiance	177	7.4
3- Moyennement confiance	379	15.9
4- Confiance	853	35.7
5- Tout à confiance	908	38.0

Représentation graphique



Remarque :

La représentation en diagramme circulaire n'est pas indiquée car on perd le caractère ordonné des modalités de la variable.

Décrire et représenter une variable numérique

Décrire une variable numérique

- Lorsque la variable est discrète (avec peu de modalités) on peut procéder comme précédemment.
- Lorsque la variable est continue on regroupe les valeurs dans des classes. La première difficulté est le choix du nombre de classes. On prendra toujours des classes de même longueur.

Parmi les solutions courantes, on peut utiliser

$$k = 1 + \frac{\log(N)}{\log(2)}$$

classes, où N est l'effectif total du jeu de données (formules de Sturges).

Exemple 2 : Variable Vuln1

on choisit par exemple des pas de 10 pour les classes

classe	effectif	fréquence (%)
[0, 10)	255	10.7
[10, 20)	415	17.4
[20, 30)	315	13.2
[30, 40)	217	9.1
[40, 50)	108	4.5
[50, 60)	243	10.2
[60, 70)	273	11.4
[70, 80)	231	9.7
[80, 90)	157	6.6
[90, 100]	175	7.3

Représenter une variable mesurée sur une échelle continue : l'histogramme

On peut représenter en ordonnées :

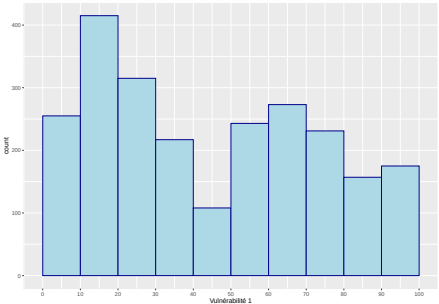
- Les effectifs de chaque classe notées n_i
- Les densités notées d_i qui sont proportionnelles aux fréquences f_i de chaque classe :

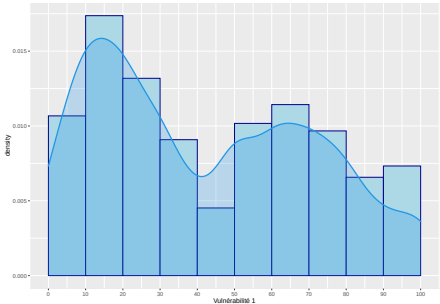
$$d_i = \frac{f_i}{l}$$

avec l la longueur de la classe (on suppose que toutes les classes ont la même longueur).

Exemple 2 : Vuln 1

Avec des intervalles de longueur 10 :





Autre choix $l = 20$

classes	fréquence	densité
[0, 10)	0.107	0.005
[10, 20)	0.174	0.009
[20, 30)	0.132	0.007
[30, 40)	0.091	0.005
[40, 50)	0.045	0.002
[50, 60)	0.102	0.005
[60, 70)	0.114	0.006
[70, 80)	0.097	0.005
[80, 90)	0.066	0.003
[90, 100]	0.073	0.004

classes	fréquence	densité
[0, 20)	0.280	0.014
[20, 40)	0.223	0.011
[40, 60)	0.147	0.007
[60, 80)	0.211	0.011
[80, 100]	0.139	0.007

Histogramme obtenu