

Document n°2. Description.

1. Décrire les variables une par une : analyse univariée.

1.1. Une variable qualitative

Les **modalités**, c'est-à-dire les valeurs possibles, d'une variable qualitative sont des **chaînes de caractères**, ou des **nombres codant ces chaînes**. Par exemple la réponse à la question « avez vous une carte de fidélité » peut être oui ou non. Il y a 2 modalités. On peut mémoriser les réponses dans le fichier par une variable composée de chaînes de caractères, « oui » ou « non », ou par une variable contenant un entier, 1 pour oui, 0 pour non.

Pour Jamovi, une variable qualitative est de **type de mesure Nominal** et de **type de données Texte** ou **Entier**.

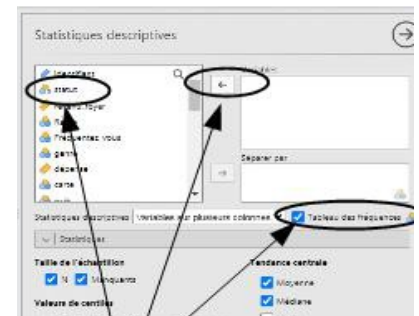
1.1.a) Traitement numérique : tri à plat.T

Il s'agit d'indiquer les différentes **modalités**, leurs effectifs – nombre d'individus présentant ces modalités– ou leurs **fréquences** – proportions de ces individus –.

Menu Analyse → Exploration → Statistiques descriptives

1. Sélectionner à gauche une variable qualitative, par exemple Statut,
2. la faire passer à droite par la flèche centrale,
3. puis cocher la case « tableau des fréquences »

Un tableau des fréquences apparaît dans la fenêtre Résultats.



Fréquences de statut

statut	Quantités	% du Total	% cumulés
Célib	20	29.4%	29.4%
En couple	39	57.4%	86.8%
Séparé	9	13.2%	100.0%

Q1. Construire un tel tableau pour trouver la proportion de femmes dans l'échantillon.

1.1.b) Représentations graphiques : diagrammes en barres, camembert.

Diagramme en barres – Bar chart – .

On représente les modalités par des rectangles dont les **aires** sont proportionnelles aux fréquences de ces modalités. Ces rectangles ont tous la même largeur, donc leurs hauteurs sont aussi proportionnelles aux fréquences de ces modalités. Les modalités ne sont pas ordonnées, donc l'ordre des rectangles n'est pas imposé.

Attention à ne pas confondre un diagramme en barres, qui décrit une variable **qualitative**, avec un **histogramme**, qui décrit une **variable quantitative continue**. Dans un histogramme, la largeur du rectangle a une signification, les rectangles sont joints les uns aux autres, ils apparaissent dans l'ordre croissant des valeurs de la variable, et il peut arriver qu'ils n'aient pas la même largeur. Dans ce cas, leurs aires sont proportionnelles aux fréquences, mais leurs hauteurs ne le sont pas.

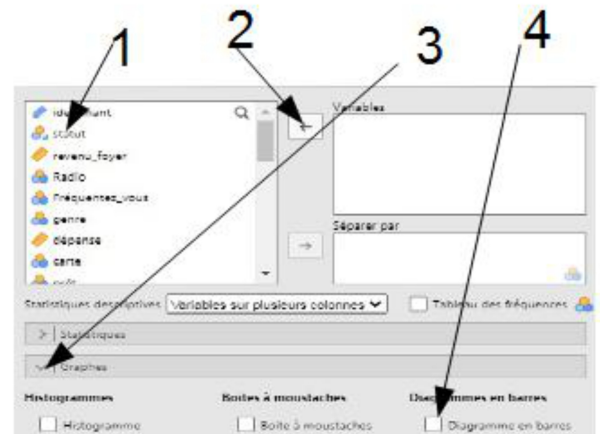
Menu **Analyse** → **Exploration** → **Statistiques descriptives**

1 Sélectionner la variable qualitative à décrire. Par exemple statut.

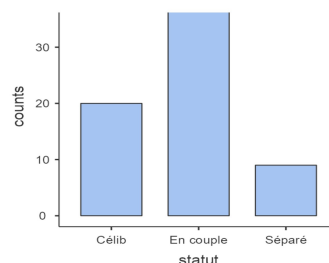
2 la transférer dans la cellule « Variables » à droite .

3 Cliquer sur **Graphes**.

4 Cocher « **Diagramme en barres** ».



Le résultat apparaît :



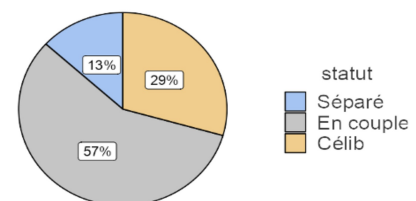
Q2. Représenter la variable Carte par un diagramme en barres

Diagrammes en secteurs, ou camemberts .

Chaque modalité est représentée par un secteur d'un disque. Les aires des secteurs sont proportionnelles aux fréquences des modalités.

Dans la barre de menus, choisir « **Jjstatsplots**», puis « **Pie Charts** »

Q3. Représenter graphiquement les modalités de Statut.



1.2. Une variable quantitative.

Elle porte sur des nombres, elle peut être discrète ou continue

Dénomination en statistiques	Signification	Exemple	Jamovi : type de données	Jamovi : type de mesure
Variable quantitative discrète	Seules quelques valeurs possibles	Taille du foyer	Décimal ou entier	ordinal
Variable quantitative continue	Un très grand nombre de valeurs différentes possibles	Revenu	Décimal ou entier	continu

1.2.a) Traitements numériques

A partir du menu Exploration → Statistiques descriptives → Statistiques , on peut choisir les statistiques parmi :

Caractéristiques de tendance centrale

- moyenne
- médiane, c'est-à-dire la valeur comparée à laquelle la moitié des observations de la variable sont inférieures, la moitié sont supérieures. On a tendance à utiliser fréquemment la moyenne mais la médiane est parfois plus adaptée. La moyenne a de plus le défaut d'être sensible aux valeurs extrêmes, qui sont souvent plus gênantes que représentatives.
- Mode, la valeur la plus fréquente. Elle est adaptée aux variables discrètes, mais en général non significative pour les variables continues. S'il y a plusieurs modes, seule une d'elles est présentée.

Caractéristiques de dispersion

- erreur standard, c'est-à-dire écart-type : $\sigma = \sqrt{\overline{(x - \bar{x}_n)^2}} = \sqrt{\frac{1}{n} \left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)}$, où x_i désigne la i ème valeur de la variable x , et $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ est la moyenne des valeurs de cette variable,
- variance $\sigma^2 = \overline{(x - \bar{x}_n)^2} = \frac{1}{n} \left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)$,
- Etendue (valeur maximale de la variable – valeur minimale de la variable).
- Ecart interquartile : $Q_3 - Q_1$, Q_3 étant la valeur atteinte ou dépassée par $\frac{1}{4}$ des individus , Q_1 la valeur atteinte ou dépassée par $\frac{3}{4}$ des individus.

Caractéristiques de forme de distribution.

- coefficient d'acuité, ou encore kurtosis . $\left(\frac{1}{n-1} \right) \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^4}{\sigma^4} - 3$ mesure l'acuité de la variable x :
plus ce coefficient est grand, plus la distribution de la variable x présente un pic marqué autour de la moyenne, et des queues de distribution élevées, c'est-à-dire des valeurs extrêmes probables.
On le compare souvent à celui d'une distribution Gaussienne, pour laquelle il vaut 0 . Une variable qui présente un kurtosis positif atteint des valeurs extrêmes plus souvent qu'une distribution Gaussienne.
Attention d'autres logiciels peuvent ne pas soustraire 3 dans la formule, dans ce cas le kurtosis est toujours positif et vaut 3 pour une distribution Gaussienne.

- Coefficient d'asymétrie. $\left(\frac{1}{n-1} \right) \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^3}{\sigma^3}$ mesure l'asymétrie de la distribution de la variable x .
Un coefficient d'asymétrie négatif est obtenu pour une distribution possédant une médiane supérieure à la moyenne, et une queue de distribution plus épaisse pour les plus petites valeurs de la variable que pour les plus grandes. Un coefficient d'asymétrie positif est obtenu pour une distribution possédant une médiane inférieure à la moyenne, et des valeurs extrêmes supérieures à la médiane plus probables que les valeurs extrêmes inférieures à la médiane.
Une distribution Gaussienne est symétrique et possède un coefficient d'asymétrie 0.

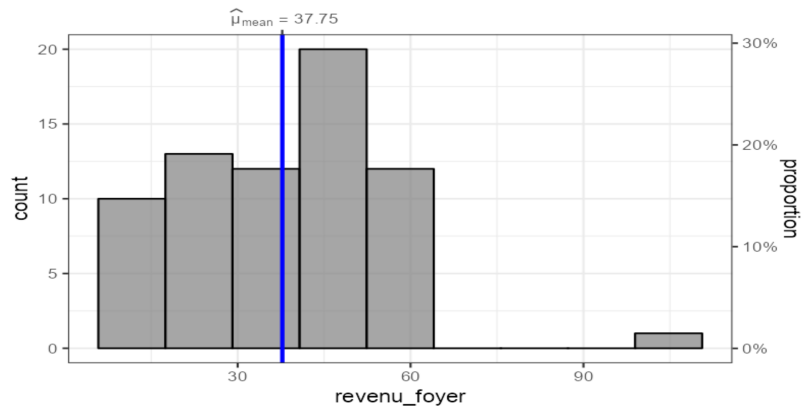
1.2.b) Traitements graphiques

Représentation graphique des fréquences : histogramme.

Il s'agit de préciser avec quels fréquences apparaissent les différents intervalles de valeurs. Dans un plan, on trace un axe décrivant les valeurs possibles de la variable, on le divise en quelques intervalles, et à la verticale de chaque intervalle on construit un rectangle représentant la fréquence d'apparition des valeurs de cet intervalle.

Les **aires** (hauteur $\times$ largeur) de ces rectangles sont **proportionnelles aux fréquences** d'apparition des intervalles : si la variable apparaît 3 fois plus souvent dans un intervalle A que dans un intervalle B, le rectangle au-dessus de A aura une aire 3 fois plus grande.

Attention 1 : si les intervalles ont la même amplitude – ce qui est le cas pour ceux que jamovi construit – alors les hauteurs des rectangles aussi sont proportionnelles aux fréquences des intervalles correspondants. Mais si les intervalles n'ont pas la même amplitude, ce n'est pas le cas ; les aires sont proportionnelles aux fréquences des intervalles, les hauteurs sont proportionnelles aux quantités suivantes : $\frac{\text{fréquence de l' intervalle}}{\text{largeur de l' intervalle}}$.



On peut obtenir un histogramme par Analyse → Exploration → Statistiques descriptives → graphes

mais surtout par **Analyse** → **JJStatsplot** → **Histogram**, qui fournit plus d'informations et permet notamment de choisir la largeur des intervalles en cochant la case « **Change bin width** »

On peut comparer les histogrammes sur des sous-populations en indiquant une variable qualitative iou une variable discrète dans la cellule « Split by »

Q4. Tracer l'histogramme de la variable `revenu_foyer`. Comment qualifier sa répartition ? Est-elle la même pour ceux qui ont ou qui n'ont pas de carte de fidélité ?

2. Décrire les interactions des variables : analyse bivariée

2.1. Variables nominales ou ordinales : tableaux de contingence

Il s'agit d'indiquer dans un tableau la distribution d'une population selon 2 ou plusieurs variables .

Cas de 2 variables nominales ou ordinales (i e quantitatives discrètes ou qualitatives).

Question 5 Sélectionner

Analyse → **fréquences** → **Tables de contingence** → **Echantillons indépendants**. Sélectionner pour variable ligne le statut et pour variable colonne le type d'émission préférée . Qu'indique le résultat ?

Tables de contingence

statut	type_emission				Total
	1	2	3	4	
Célib	13	3	3	1	20
En couple	12	13	14	0	39
Séparé	3	2	4	0	9
Total	28	18	21	1	68

Un menu déroulant « Cellules » permet de faire apparaître le pourcentage de population de chaque cellule par rapport à toute la ligne, toute la colonne, ou toute la population. Il permet aussi de faire apparaître la « population attendue » en cas d'indépendance des deux variables, ce qui sera utile pour mener un test du Chi2 d'indépendance.

En dessous du tableau apparaît le résultat d'un test de Chi2 d'indépendance, que nous verrons à la partie 3 du cours, et qui permet de juger si les variables paraissent raisonnablement indépendantes.

Enfin, on peut indiquer dans la cellule « **quantités** », une variable quantitative. Le tableau contiendra alors au lieu des effectifs – le nombre d'individus concernés – de chaque case, **la somme pour ces individus de la variable quantitative** choisie. Par exemple, la dépense cumulée des individus concernés.

2.2. Variables quantitatives : calcul du coefficient de corrélation

Le coefficient de corrélation entre 2 variables X, Y, est $\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}} = \frac{\overline{X \times Y} - (\overline{X})(\overline{Y})}{(\sqrt{\overline{X^2} - (\overline{X})^2})(\sqrt{\overline{Y^2} - (\overline{Y})^2})}$

Il est toujours compris entre -1 et 1, il est d'autant plus proche de 1 que X et Y ont tendance à varier, en moyenne, dans le même sens, d'autant plus proche de -1 que X et Y ont tendance à varier, en moyenne, en sens inverse.

Lorsqu'on calcule les coefficients de corrélation entre plusieurs variables, il est d'usage de les présenter dans une matrice : la matrice de corrélation.

Chaque ligne et chaque colonne correspond à une des variables choisies. A l'intersection de la ligne de la variable X et de la colonne de la variable Y se trouve le coefficient de corrélation entre X et Y.

Entre n variables on peut calculer n² coefficients de corrélation, mais il faut remarquer que $\rho_{X,Y} = \rho_{Y,X}$ et $\rho_{X,X} = 1$. Il est donc inutile de préciser $\rho_{X,X}$, et si l'on connaît $\rho_{X,Y}$ il est inutile de préciser $\rho_{Y,X}$. C'est pourquoi Jamovi n'affiche que les coefficients de corrélation situés sous la diagonale de la matrice de corrélation.

Voici un exemple qui pourrait être fourni par Jamovi :

$$\begin{pmatrix} 0,5 & & \\ \mathbf{0,1} & -0,3 & \\ 0,2 & 0,9 & -0,4 \end{pmatrix} \quad \text{Le coefficient de corrélation}$$

entre les variables n°1 et n°3 est 0,1, le même se trouverait à la 1ère ligne, 3è colonne si l'on remplissait la matrice. D'autre part sur la diagonale il n'y a que des 1, qui ne sont pas représentés.

On peut calculer ces coefficients par **Analyse → régression → matrice de corrélation.**

Question 6 Calculer les coefficients de corrélation entre la dépense, le revenu, la taille du foyer, l'année de naissance. Qu'en penser ?

Matrice de corrélation				
	revenu_foyer	dépense	taille_foyer	naissance
revenu_foyer	—			
dépense	0.281	—		
taille_foyer	0.636	0.339	—	
naissance	-0.146	0.110	-0.039	—

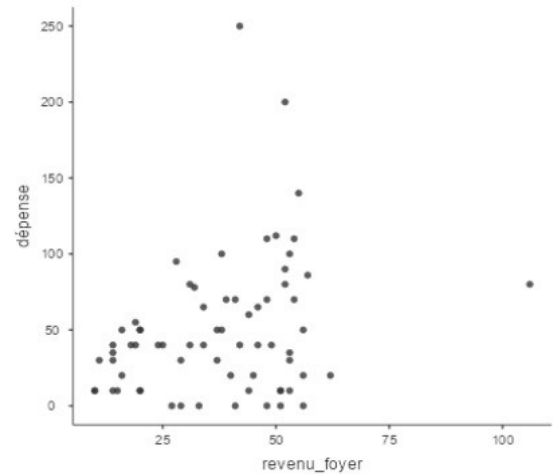
2.3. Représentation graphique de la relation entre deux variables.

2.3.a) Relations entre 2 variables quantitatives : nuage de points.

On l'obtient par **Analyse** → **Exploration** → **nuage de points**.

Le fenêtre qui s'ouvre propose de choisir les deux variables quantitatives portées en abscisses et en ordonnées.

On peut indiquer une troisième variable, qualitative, qui définit des sous-populations. Pour chacune d'elles sera utilisée une couleur différente des autres. Malheureusement on ne peut pas choisir les couleurs utilisées, le résultat est parfois peu lisible...



Question 7 Représenter la dépense (en ordonnées) en fonction du revenu du foyer (en abscisses).

2.3.b) relations entre 2 variables qualitatives – ordinales ou nominales – .

Diagramme en barres d'une variable qualitative en fonction des valeurs d'une autre variable.

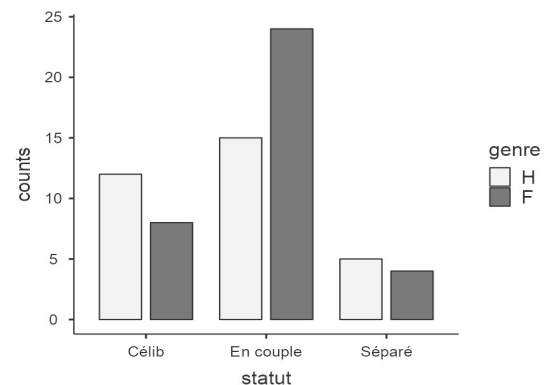
Construisons un diagramme en barres représentant les fréquences des trois modalités de la **variable Statut** : « Célibataire », « En couple », « Séparé », mais en détaillant pour chaque statut la **répartition de genre**.

Question 8

Utiliser **Analyse** → **Exploration** → **Statistiques descriptives** → **Graphes** . Cocher la case « **Diagrammes en barres** ».

Revenir en haut de la fenêtre, choisir la variable **Statut** dans la cellule **Variables** , et **Genre** dans la cellule **Séparer par**.

La répartition des sexes est-elle semblable pour tous les statuts ?



2.3.c) Relations entre 1 variable quantitative et une variable qualitative : boîtes de dispersions et histogrammes comparés.

Boîte à moustaches – boîte de dispersion, boxplot – d'une variable quantitative selon les valeurs d'une variable qualitative.

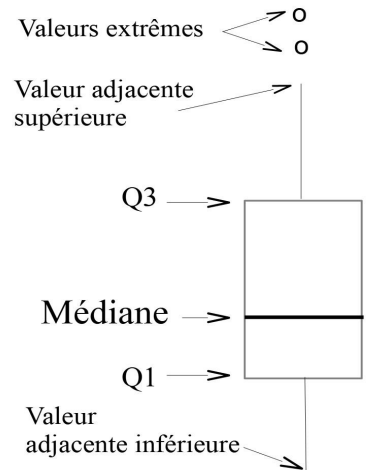
Elle représente la tendance centrale, la dispersion, et les valeurs extrêmes d'une variable quantitative.

Les côtés haut et bas du rectangle sont aux ordonnées Q1 , 1er quartile, valeur dépassant 25% des observations, et Q3, 3ème quantile, valeur dépassant 75% des observations.

A l'intérieur, un segment est situé à la médiane.

La moustache supérieure est la valeur adjacente supérieure : la plus haute valeur qui reste inférieure à $Q3 + 1,5 (Q3 - Q1)$, la moustache inférieure est la valeur adjacente inférieure : la plus basse valeur qui reste supérieure à $Q1 - 1,5 (Q3 - Q1)$

A l'intérieur de la boîte se trouvent la moitié des observations. Si la variable suit une distribution Gaussienne, environ 99% des observations se trouvent entre les deux moustaches. Les valeurs hors de cet intervalle sont atypiques et il faut s'interroger sur leur présence : n'y a-t-il pas d'erreur de saisie ? Les observations ont-elles leur place dans l'échantillon ? Qu'est-ce qui les différencie des autres ?



Une boîte seule a peu d'intérêt, en revanche **plusieurs boîtes** permettent de **comparer** la distribution dans différents groupes d'observations.

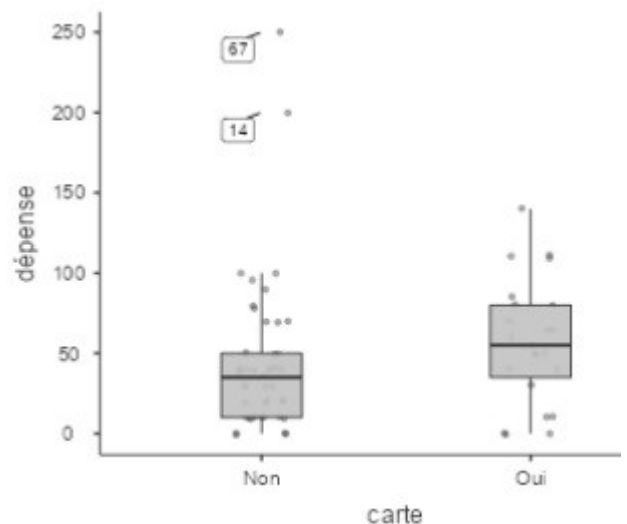
Question 9 Tracer les boîtes de dispersion de dépense (variable quantitative continue) avec ou sans carte de fidélité (donc selon les valeurs de la variable qualitative carte)

Utiliser : **Analyse** → **Exploration** → **Statistiques descriptives** → **Graphes** . Cocher « **Boîtes à moustaches** ».

Dans la cellule **Variables** , indiquer la variable **dépense** et dans la cellule **Séparer par** indiquer **carte**.

Jamovi permet aussi d'identifier les valeurs extrêmes en cochant « Etiqueter les valeurs aberrantes » , de représenter les données par des points en cochant « données » , et de représenter plus finement la distribution de ces points en cochant « violon » . qui trace un violon autour de la boîte de dispersion, la largeur du violon à une ordonnée représentant la densité de valeurs à cette ordonnée.

Que peut-on en déduire sur la dépense comparée des client ayant ou n'ayant pas de carte de fidélité ?



Histogramme d'une variable quantitative selon les valeurs d'une variable qualitative.

On peut aussi juger de l'influence d'une variable qualitative sur une variable quantitative continue en comparant des histogrammes de cette variable continue.

Question 10 Comme au paragraphe 1.2.b) , réaliser un histogramme par **Analyse** → **Exploration** → **Statistiques descriptives** → **Graphes** et en cochant « **Histogramme** » . Dans la cellule Variables indiquer la variable continue **Dépense** , et dans la cellule « **Séparer par** », indiquer la variable qualitative **Carte**.

Différents histogrammes, selon les valeurs de la variable quantitative, sont accolés, pour qu'on puisse comparer les distributions de la variable quantitative continue. **Commenter ce graphique.**

