Séance 2 (2) : statistique descriptive bivariée

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

1. Deux variables catégorielles
2. Deux variables numériques
3. Catégoriel X numérique

Section 1

Deux variables catégorielles

Vous avez dit “bivarié” ?

- On considère **conjointement deux variables X et Y** pour
  - analyser les valeurs prises par chacune des deux variables
  - étudier le lien éventuel entre les deux variables (corrélation)
- Exemples : genre et réussite au baccalauréat ; fréquentation des cinémas et niveau d’études.

Nous allons principalement nous concentrer sur un outil essentiel de la statistique bivariée : le **tableau de contingence**.

Sous-section 1

Le tableau de contingence

Le tableau de contigence

Aussi appelé “tableau croisé” ou “tableau à double entrée”.

> *Si les statisticiens préfèrent, en toute rigueur, parler de « table de contingence », les sociologues utilisent plus facilement l'expression de « tableau croisé », plus imagée, pour désigner cet outil qui à lui seul incarne, en même temps qu'il la symbolise, toute une façon de faire de la sociologie.*

Pierre Mercklé - Les 100 mots de la sociologie

La table de contingence

- Les lignes correspondent aux valeurs (discrètes, classes ou modalités) de la variable X, les colonnes à celles de Y.
- Chaque case contient l'effectif des individus pour lesquels les deux variables prennent les valeurs correspondant à celles de la ligne et de la colonne.

	X_1	X_2	X_3
Y_1			
Y_2			
Y_3	Eff. $X = X_1$ et $Y = Y_1$		

Notations du cas général

- Variables X (k modalités : $\{x_1, ..., x_k\}$), Y (l modalités : $\{y_1, ..., y_l\}$) : **« tableau à l lignes et k colonnes »**
- $n_{i,j}$: **« nombre d'individus pour lesquels $X = x_j$ ET $Y = y_i$ »** *ou* **« case de la i-ième ligne et j-ième colonne »**

/	$X = x_1$	$X = x_2$	...	$X = x_j$	...	$X = x_k$
$Y = y_1$	$n_{1,1}$	$n_{1,2}$	...	$n_{1,j}$	...	$n_{1,k}$
$Y = y_2$	$n_{2,1}$	$n_{2,2}$	...	$n_{2,j}$	...	$n_{2,k}$
...	...	...	...	...	...	
$Y = y_i$	$n_{i,1}$	$n_{i,2}$	...	$n_{i,j}$	...	$n_{i,k}$
...	...	...	...	...	...	
$Y = y_l$	$n_{l,1}$	$n_{l,2}$	...	$n_{l,j}$	...	$n_{l,k}$

Exercice de lecture

On s'intéresse aux séries des 331994 personnes ayant passé le baccalauréat général en 2013 (chiffres de l'inscription, et non de réussit). On présente les résultats sous forme d'un tableau croisé.

	Fille	Garçon
L	43735	11589
S	77460	92409
SES	64721	42080

Effectif total

La somme des effectifs dans chaque case doit être égale au nombre total d'individus n (effectif total).

	Fille	Garçon
L	43735	11589
S	77460	92409
SES	64721	42080

[1] 331994

Cas général :

$$\sum_{i=1}^{l}\sum_{j=1}^{k} n_{i,j} = n$$

Marges

Par convention, les **effectifs lignes** sont ajoutés au tableau dans une colonne supplémentaire, et les **effectifs colonnes** dans une ligne supplémentaire.

	Fille	Garçon	Ensemble
L	43735	11589	55324
S	77460	92409	169869
SES	64721	42080	106801
Ensemble	185916	146078	331994

Notation du cas général

/	$X = x_1$	...	$X = x_j$	...	$X = x_k$	Ensemble
$Y = y_1$	$n_{1,1}$	...	$n_{1,j}$	...	$n_{1,k}$	$n_{1\cdot}$
...	...	...	...	...	...	
$Y = y_i$	$n_{i,1}$	...	$n_{i,j}$	...	$n_{i,k}$	$n_{i\cdot}$
...	...	...	...	...		
$Y = y_l$	$n_{l,1}$	...	$n_{l,j}$	...	$n_{l,k}$	$n_{l\cdot}$
Ensemble	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot k}$	n

Total de la i-ième ligne : $n_{i\cdot} = \sum_{j=1}^{k} n_{i,j}$

Total de la j-ième colonne : $n_{\cdot j} = \sum_{i=1}^{l} n_{i,j}$

Propriété des marges : exercice

Rappel du tableau :

	Fille	Garçon	Ensemble
L	43735	11589	55324
S	77460	92409	169869
SES	64721	42080	106801
Ensemble	185916	146078	??

- Qu'obtient-on si on fait le total des effectifs lignes ?
- Et le total des effectifs colonnes ?
- Que remarque-t-on ?

Propriété des marges : solution

Rappel du tableau :

	Fille	Garçon	Ensemble
L	43735	11589	55324
S	77460	92409	169869
SES	64721	42080	106801
Ensemble	185916	146078	??

Le total des effectifs lignes et colonnes est égal à l'effectif total.

Ici, la somme des ensembles est égale à 331994.

Fréquence totale : cas général

- À partir des effectifs

$$\frac{n_{i,j}}{n}$$

- À partir des effectifs lignes

$$\frac{n_{i.}}{n} = \frac{1}{n} \sum_{j=1}^{k} n_{i,j}$$

- À partir des effectifs colonnes

$$\frac{n_{.j}}{n} = \frac{1}{n} \sum_{i=1}^{l} n_{i,j}$$

Tableau en fréquence de l'effectif total

Le tableau final :

	Fille	Garçon	Ensemble
L	0.13	0.03	0.17
S	0.23	0.28	0.51
SES	0.19	0.13	0.32
Ensemble	0.56	0.44	1.00

Que remarque-t-on :

- En additionnant les fréquences ?
- En additionnant les fréquences marginales ?

Fréquence totale : propriété

- La somme des fréquence vaut 1 (100%).

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{n_{i,j}}{n} = \frac{n}{n} = 1$$

- La somme des fréquences marginales vaut 1 (100%).

Exemple pour les effectifs lignes :

$$\sum_{i=1}^{l}\frac{1}{n}n_{i\cdot} = \sum_{i=1}^{l}\frac{1}{n}\sum_{j=1}^{k}n_{i,j} = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{n_{i,j}}{n} = \frac{n}{n} = 1$$

Fréquences en ligne et en colonne

Il est souvent utile d'utiliser des **fréquences par rapport aux modalités des variables utilisées, et non par rapport à l'échantillon total**.

Dans un tableau de contingence, il est possible de donner les résultats en fréquences en lignes (resp. en fréquence en colonnes). Par convention, on ajoute une colonne (resp. une ligne) pour y afficher le total de chaque ligne (resp. colonne), qui doit être égal à 100%.

Attention ! Lors de la lecture d'un tableau croisé, **il ne faut pas confondre fréquences en ligne et en colonne**, au risque de commettre de grandes erreurs d'interprétation.

Fréquences marginales : ligne

Fréquences en ligne

	Fille	Garçon	Ensemble
L	0.79	0.21	1
S	0.46	0.54	1
SES	0.61	0.39	1
Ensemble	0.56	0.44	1

Fréquences marginales : colonne

Fréquences en colonne

	Fille	Garçon	Ensemble
L	0.24	0.08	0.17
S	0.42	0.63	0.51
SES	0.35	0.29	0.32
Ensemble	1.00	1.00	1.00

Fréquences marginales : cas général

Dans le cas des fréquences en ligne :

$$\frac{\text{Effectif de la case}}{\text{Effectif ligne}} = \frac{n_{i,j}}{n_{i.}}$$

Dans le cas des fréquence en colonne :

$$\frac{\text{Effectif de la case}}{\text{Effectif colonne}} = \frac{n_{i,j}}{n_{.j}}$$

Vérification du total des fréquences en ligne :

$$\sum_{j=1}^{k} \frac{n_{i,j}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1$$

Sous-section 2

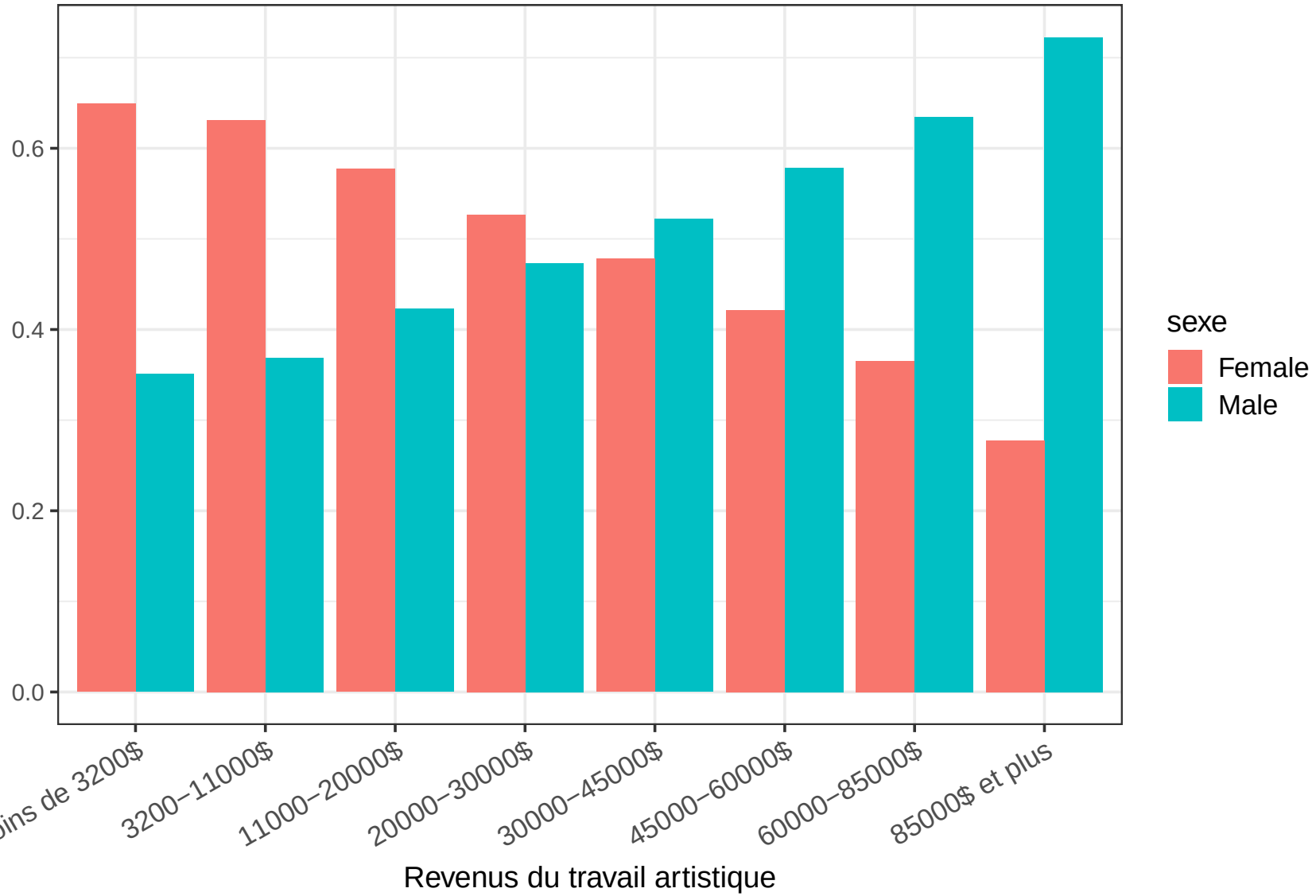
Représentation graphique

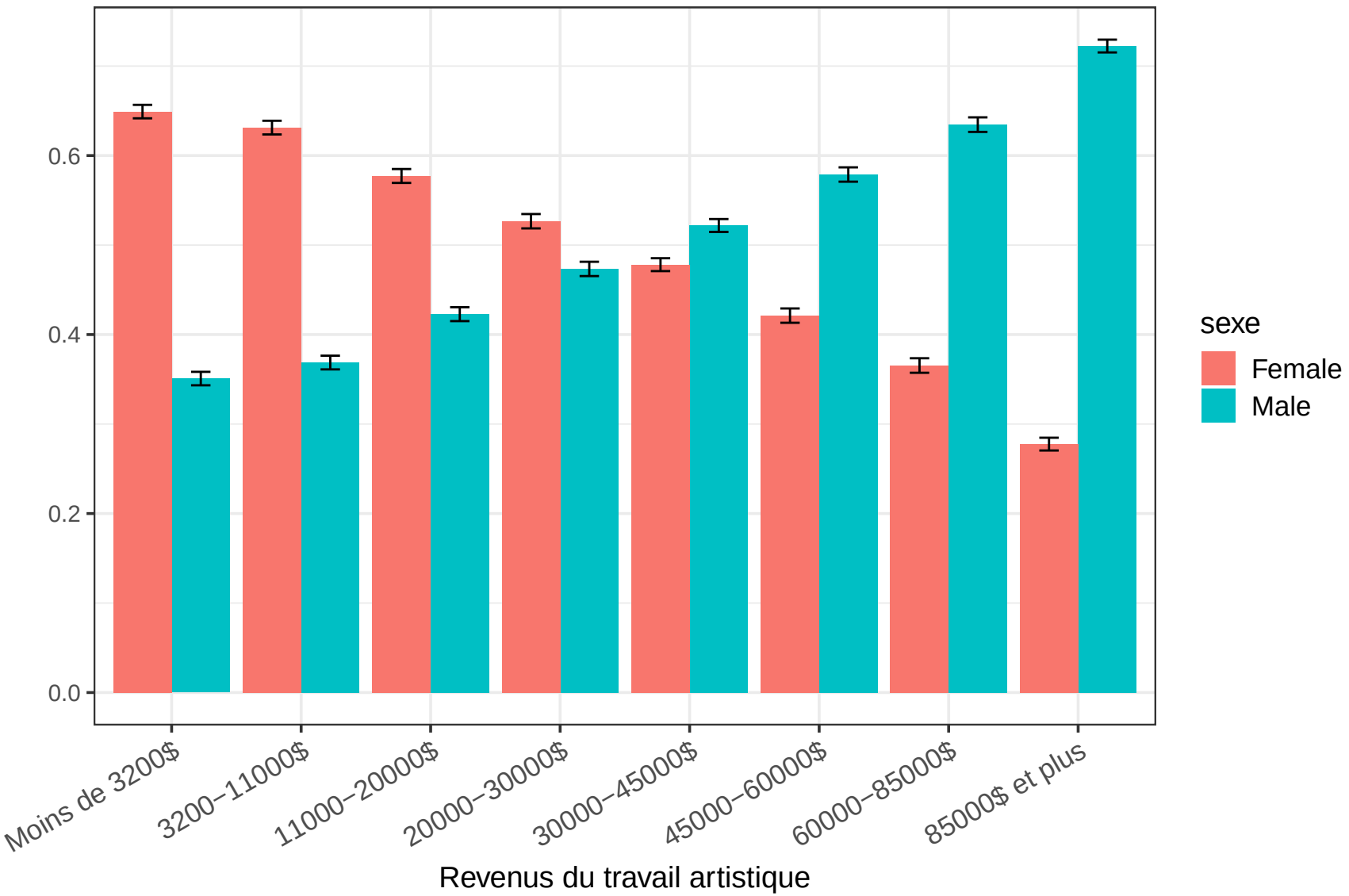
Diagramme en barre

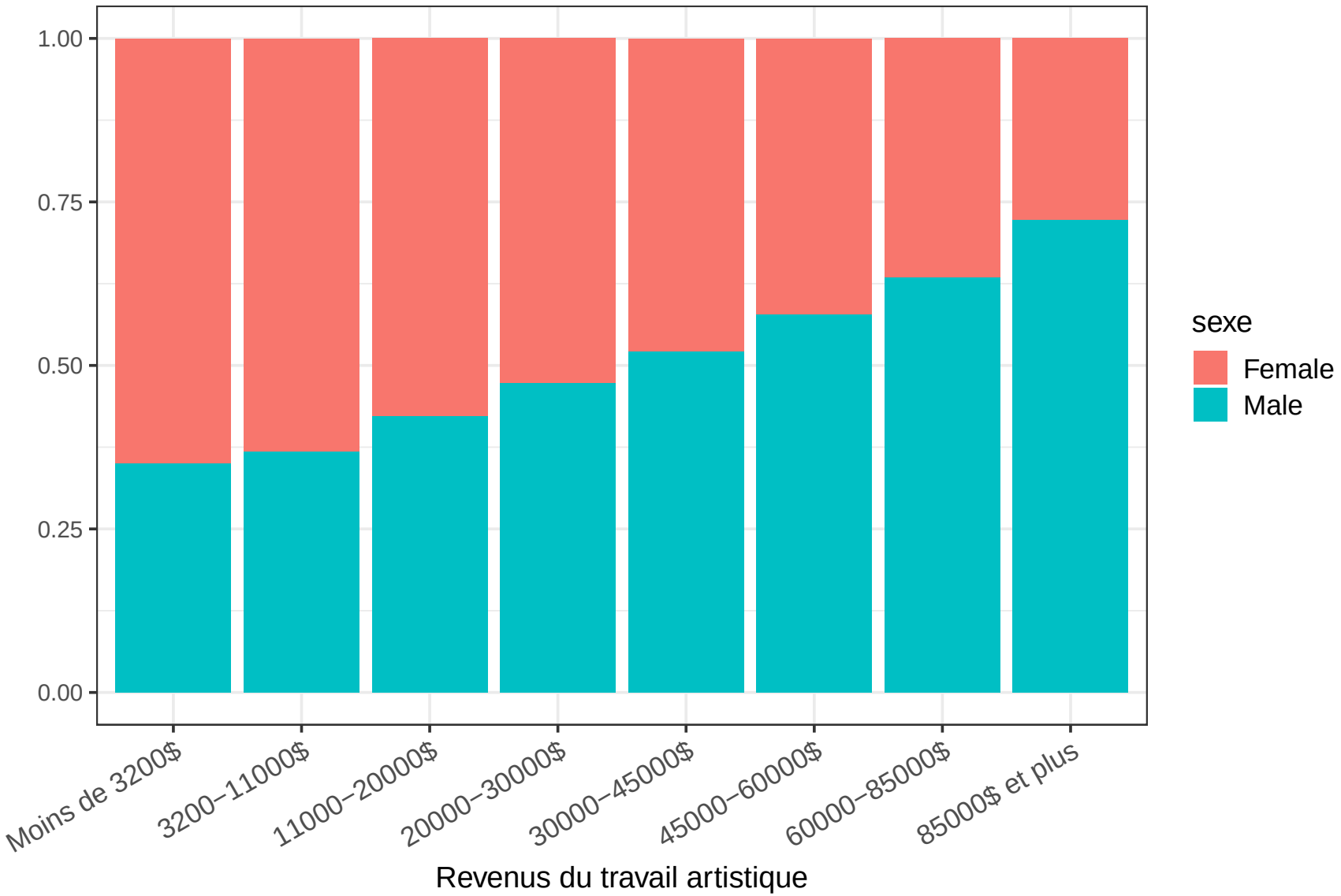
On peut utiliser le diagramme en barre pour représenter l'interaction de deux variables qualitatives. Les modalités de l'une sont représentées en abscisse. Il y a ensuite deux possibilités :

- soit on représente une barre pour chacune des modalités de la seconde variable
- soit on représente une seule barre découpée en aires différenciées par une couleur ou une texture

L'axe des ordonnées donne alors l'effectif ou la fréquence.


0.6
0.4
0.2
0.0
Moins de 3200$
3200−11000$
11000−20000$
20000−30000$
30000−45000$
45000−60000$
60000−85000$
85000$ et plus
Revenus du travail artistique
sexe
Female
Male


0.6
0.4
0.2
0.0
Moins de 3200$
3200−11000$
11000−20000$
20000−30000$
30000−45000$
45000−60000$
60000−85000$
85000$ et plus
Revenus du travail artistique
sexe
Female
Male


1.00
0.75
0.50
0.25
0.00
Moins de 3200$
3200−11000$
11000−20000$
20000−30000$
30000−45000$
45000−60000$
60000−85000$
85000$ et plus
Revenus du travail artistique
sexe
Female
Male

Section 2

Deux variables numériques

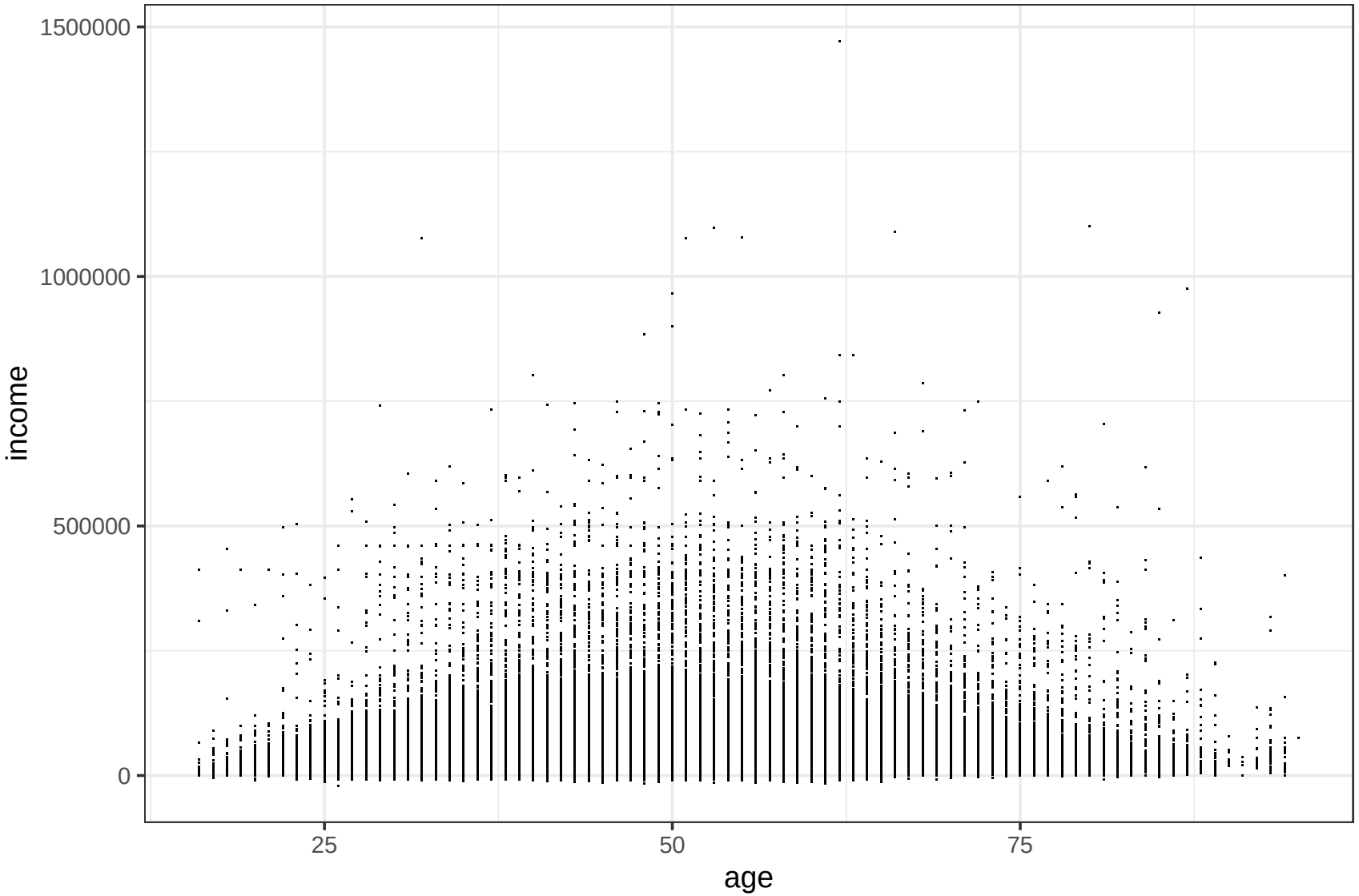
Sous-section 1

Représentation graphique

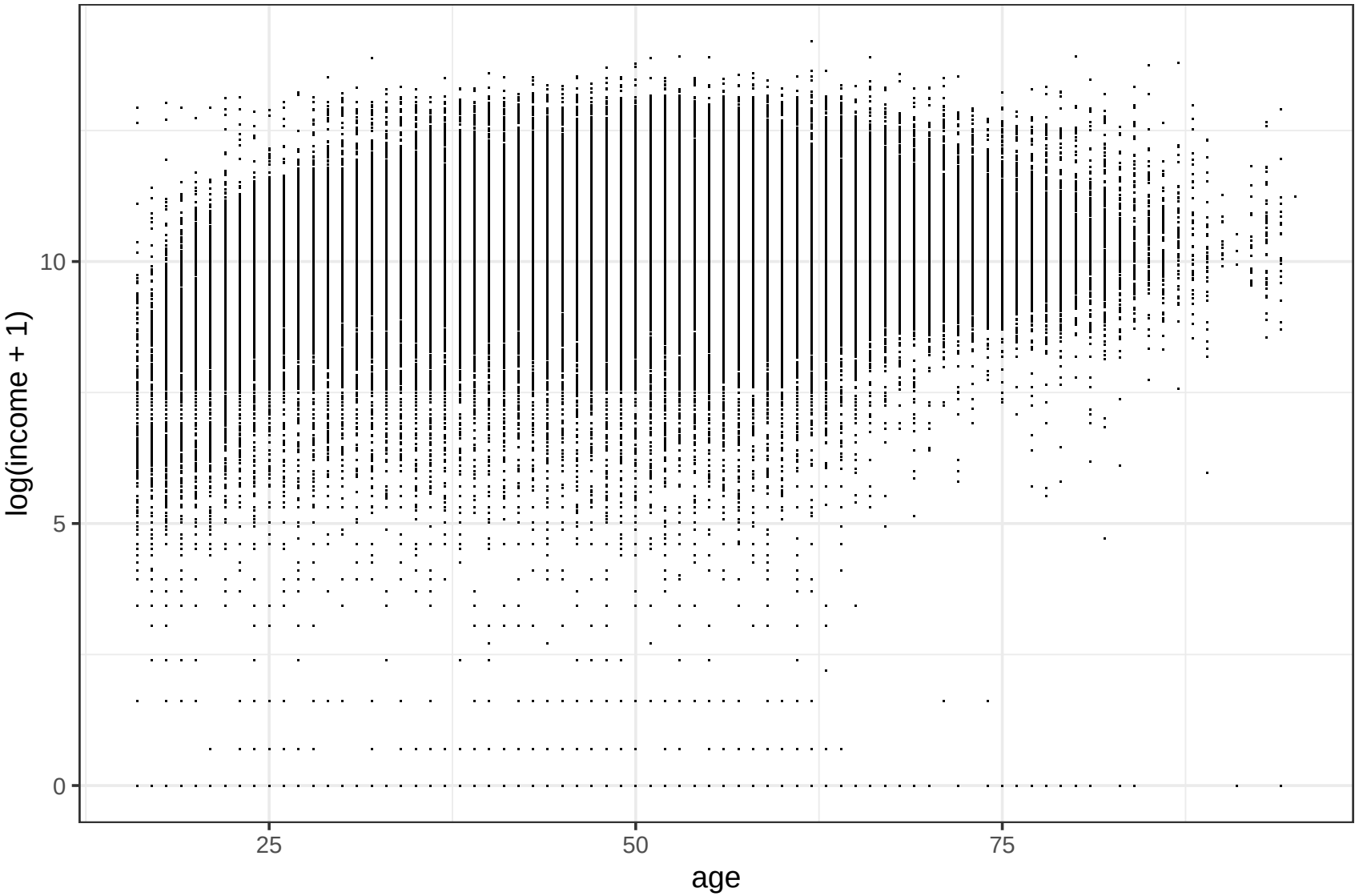
Représentation graphique

La représentation graphique la plus simple et la plus utile pour représenter deux variables numériques est le diagramme en point (*scatterplot*). Chaque observation est projetée sur un plan avec pour coordonnées la valeur de chacune des deux variables représentées.

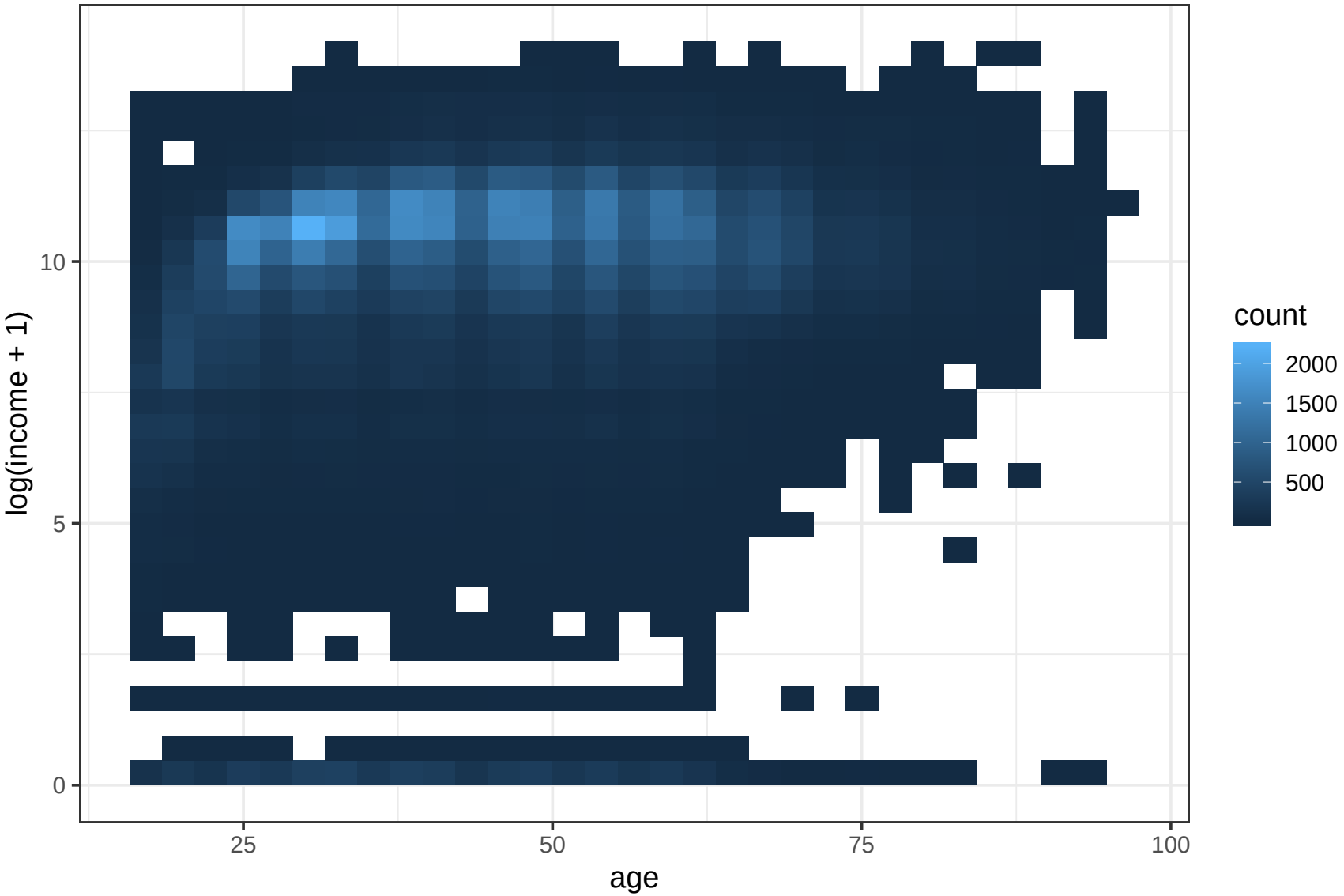
Age et revenu



Age et log du revenu

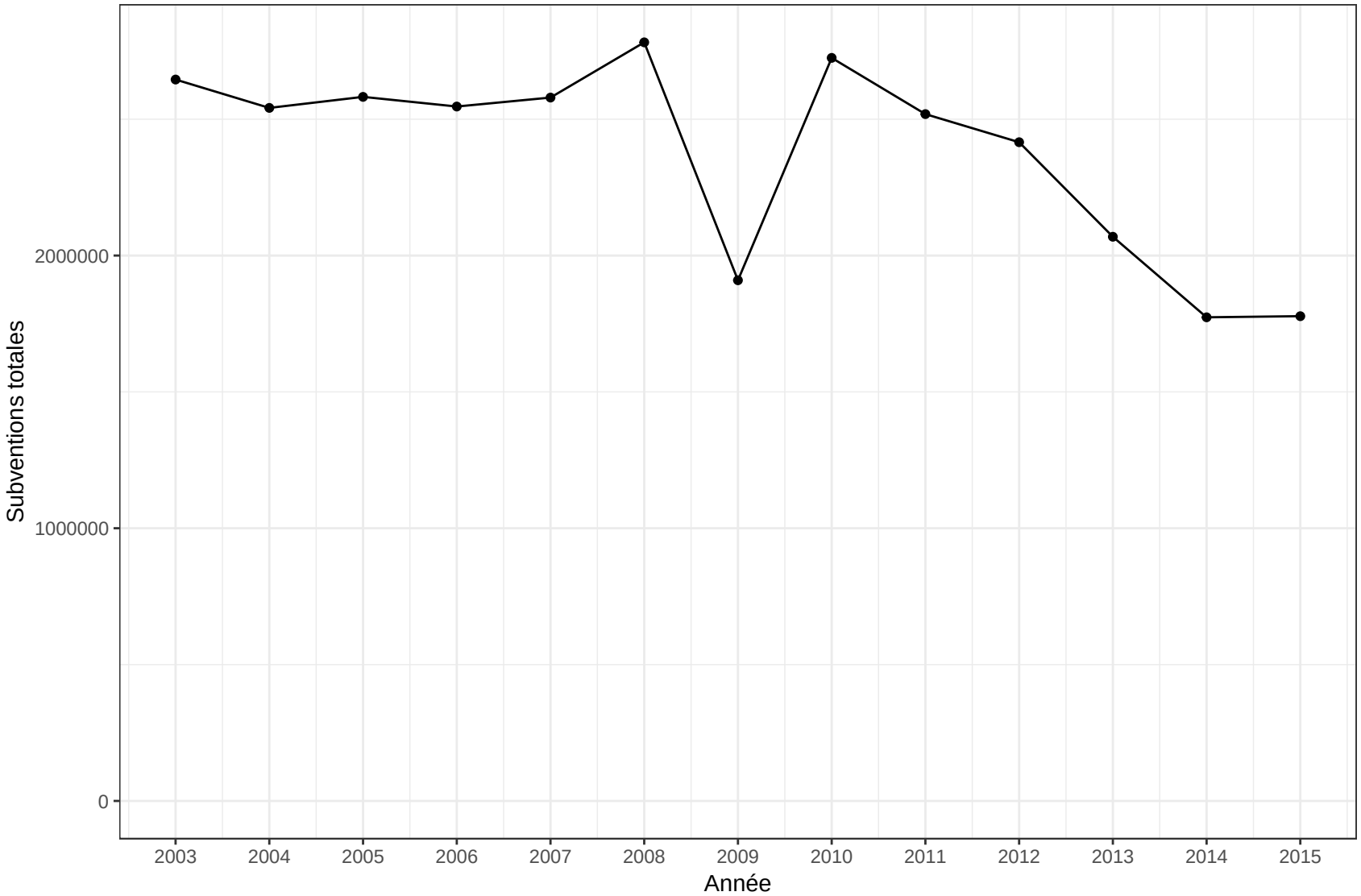


Age et log du revenu : densité

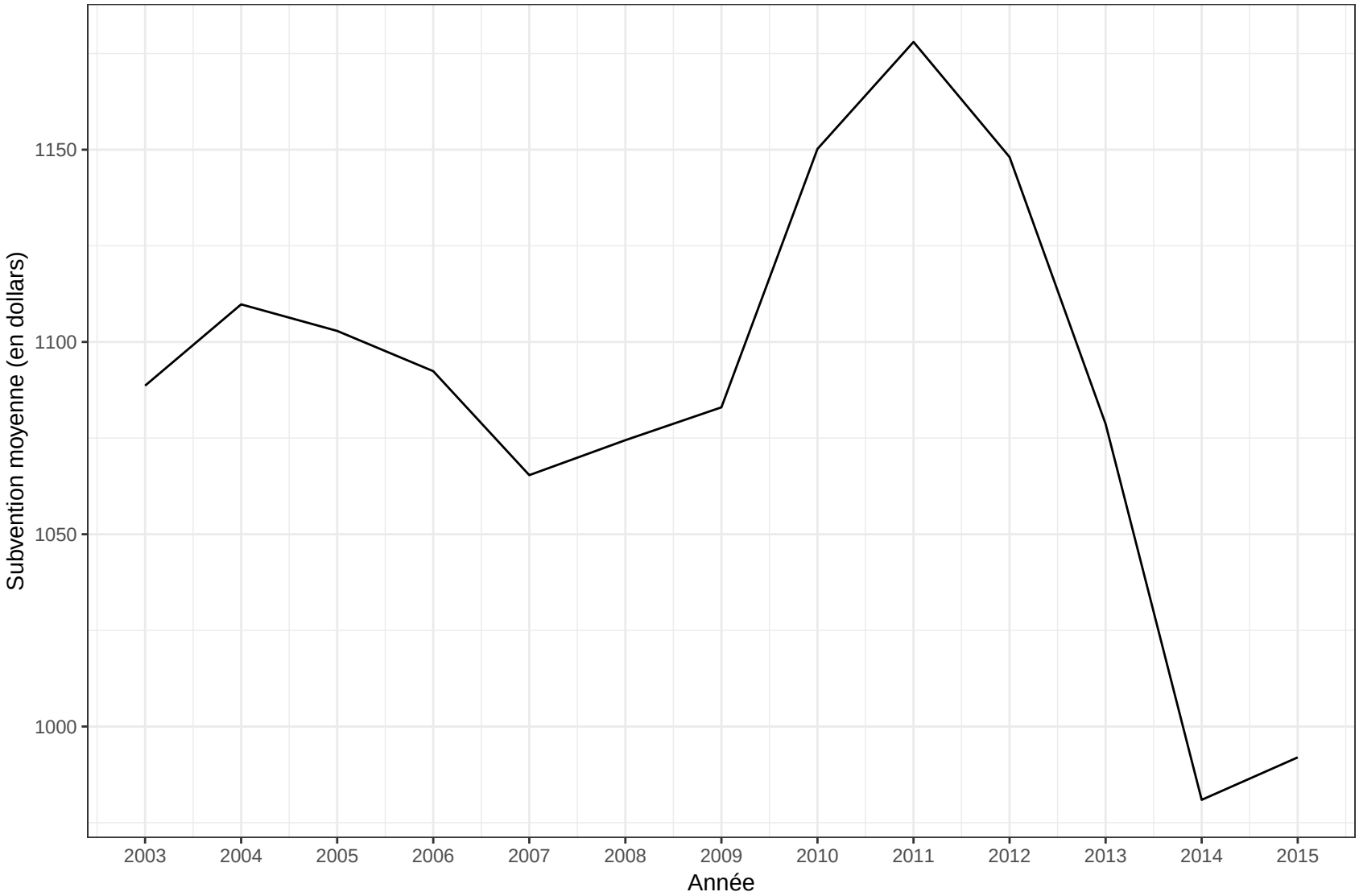


Séries temporelles

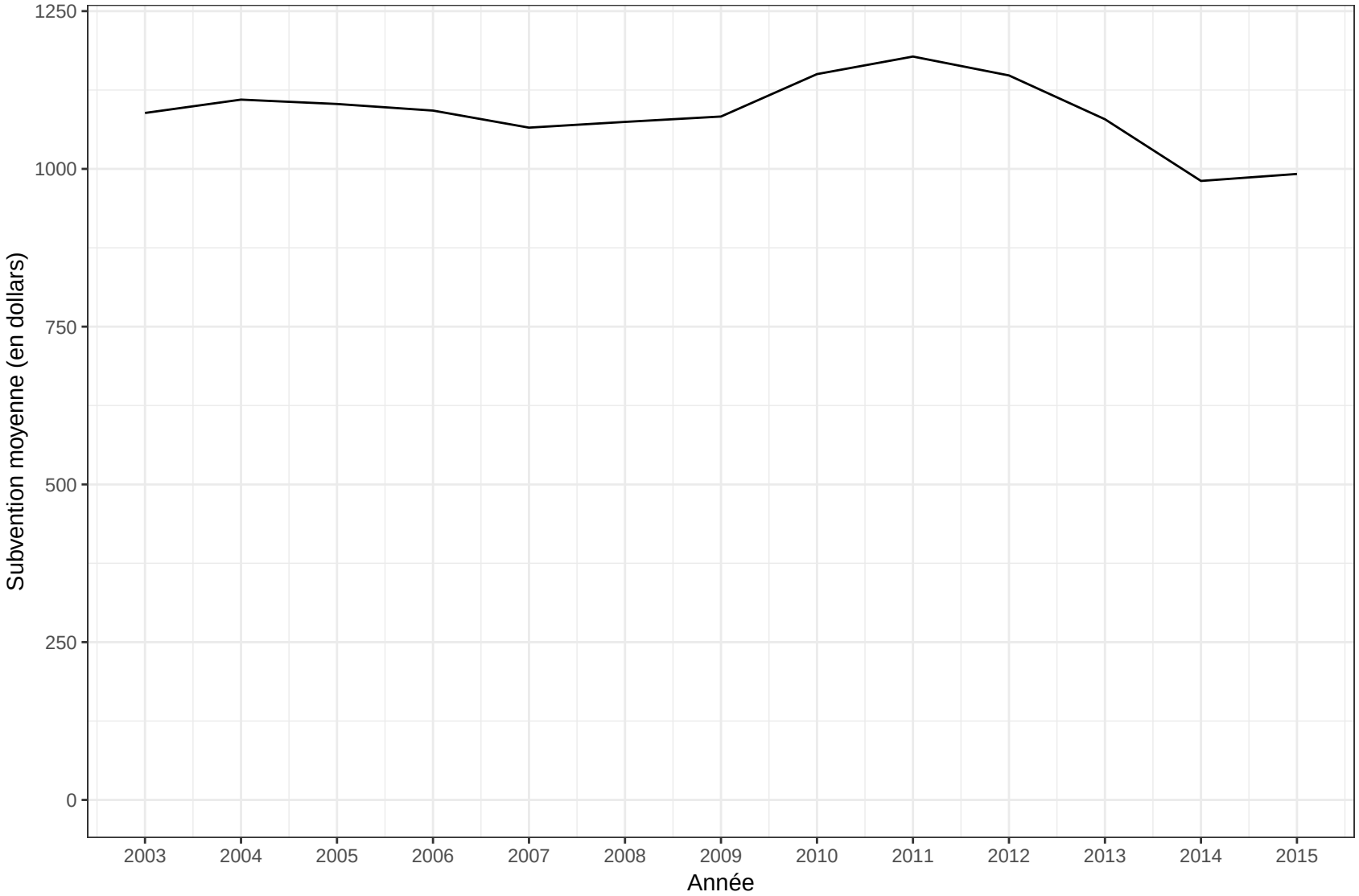
Une série temporelle représente l'évolution d'une ou de plusieurs valeurs sur une longue période. On réserve habituellement le diagramme en ligne à de telles séries continues.


Subventions totales
2000000
1000000
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Année

Moyenne des subventions par an



Moyenne des subventions par an (origine à 0)



Sous-section 2

Indicateurs

Le coefficient de corrélation

Le coefficient r de Pearson permet de mesurer le degré de corrélation linéaire entre deux variables quantitatives X et Y. Il est compris entre [-1 ; 1] :

- Le signe indique une corrélation positive ou négative
- 0 signifie une absence de corrélation
- Plus la valeur absolue est proche de 1, plus le degré de corrélation linéaire est élevé

Il se calcule de la manière suivante :

$$r = \frac{cov(X, Y)}{\sigma(X)\sigma(Y)}$$

La covariance

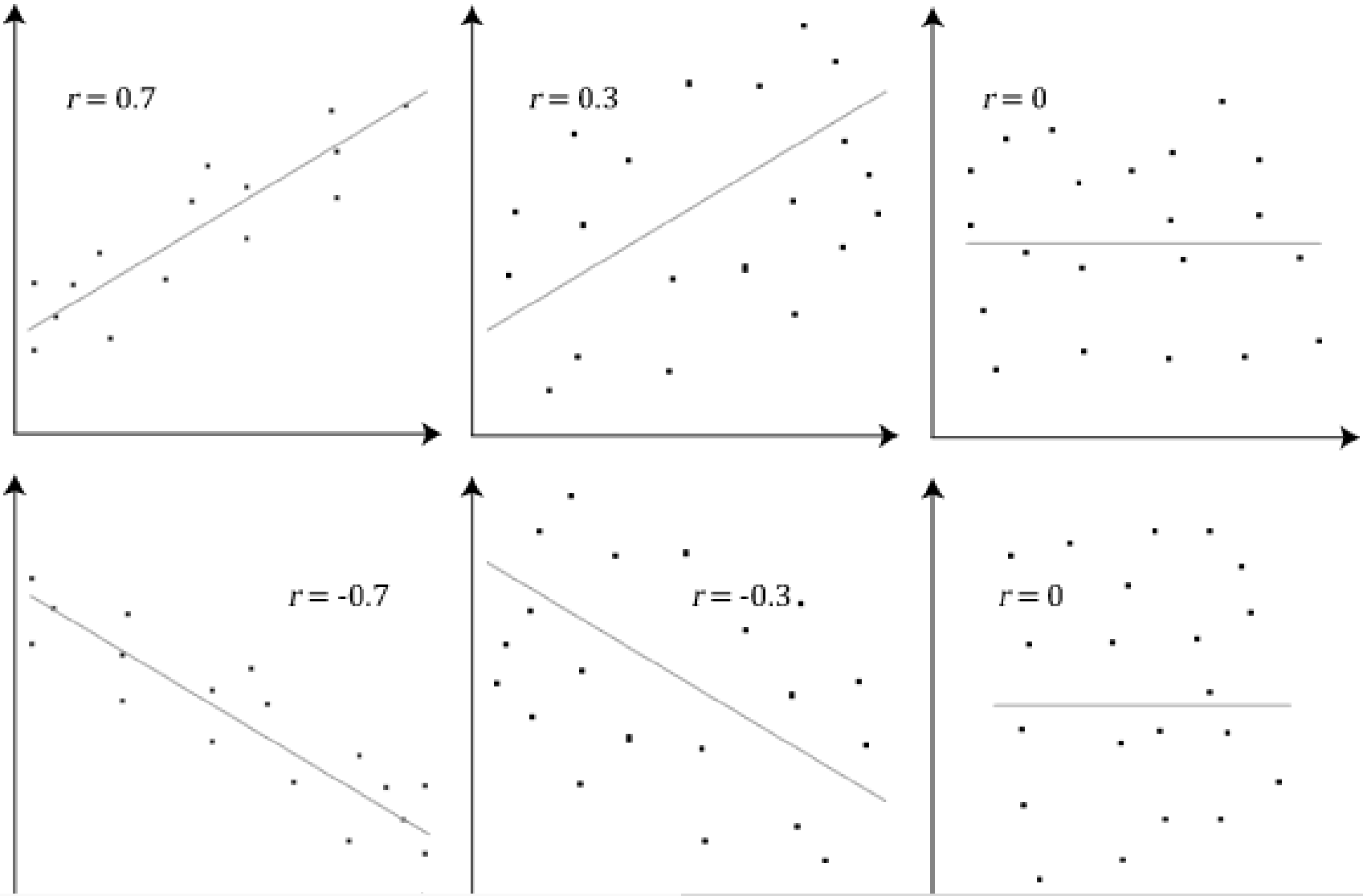
La covariance de deux variables X et Y permet d'estimer dans quelle mesure deux variables changent conjointement.

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})$$

Le signe de la covariance nous permet d'estimer si le lien entre les deux variables est positif ou négatif, mais l'ampleur de la covariation n'est pas simple à estimer (parallèle avec la variance).

C'est pourquoi le r de Pearson normalise la covariance avec le produit des écarts-types.

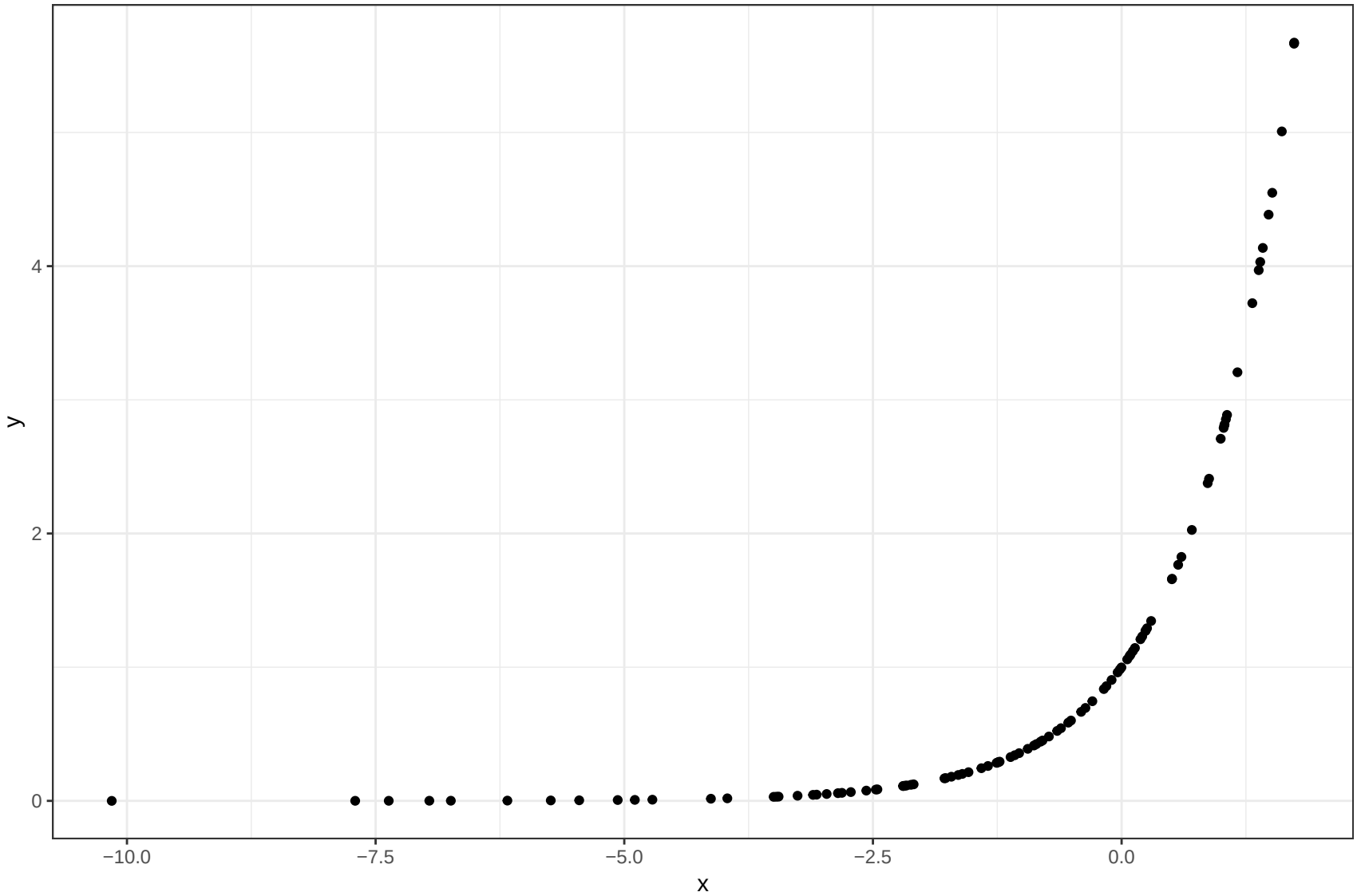
r de Pearson : lecture graphique



La corrélation linéaire

Le r de Pearson mesure seulement la force de la *corrélation linéaire* entre deux variables. D'autres relations, non-linéaires, peuvent exister. Sur la figure suivante, les deux variables sont parfaitement corrélées ($x = log(y)$), mais le r de pearson est seulement de 0.6865541 car la relation n'est pas linéaire ($x = ay + b$).

Corrélation logarithmique



Section 3

Catégoriel X numérique

Sous-section 1

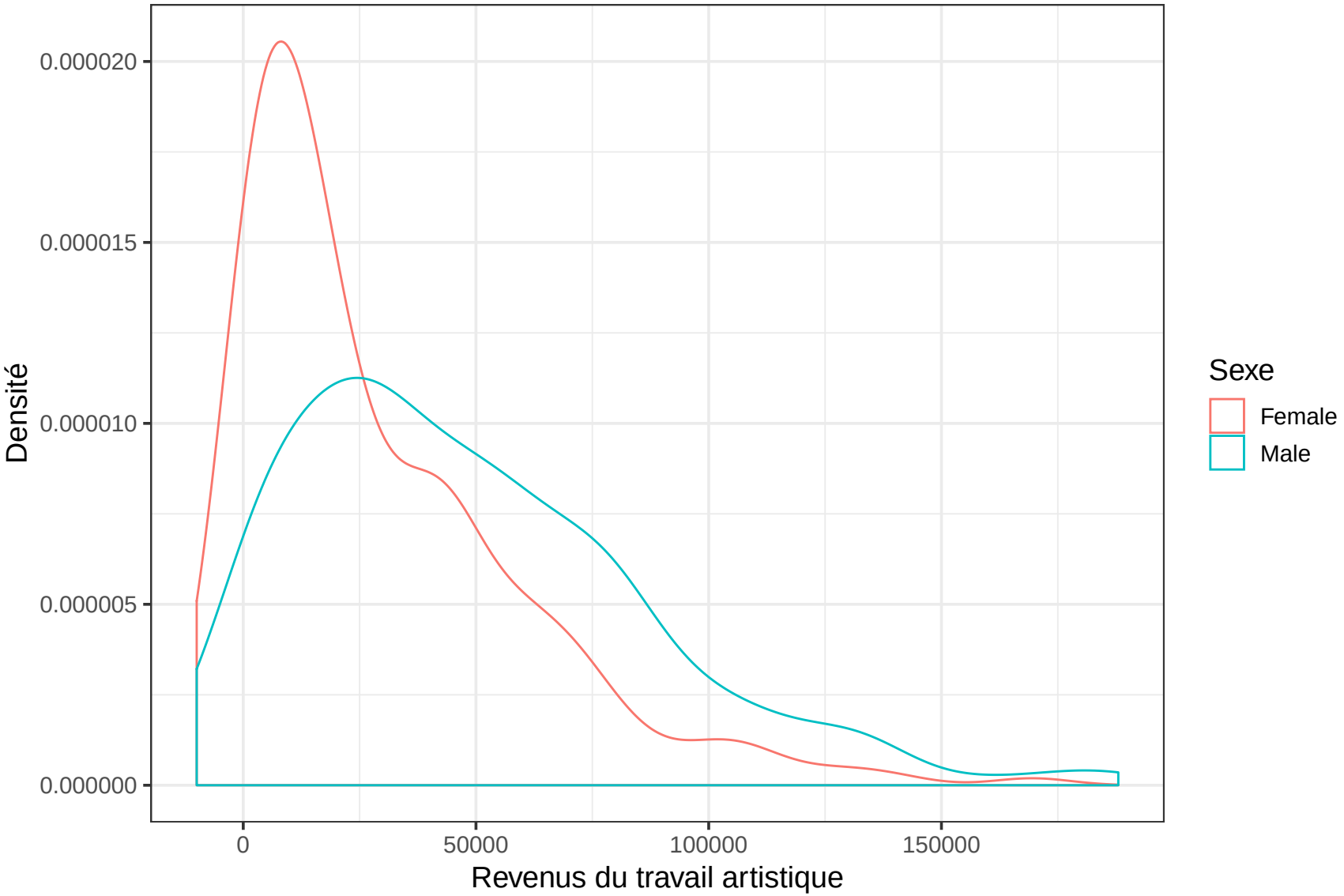
Représentation graphique

Représentation graphique

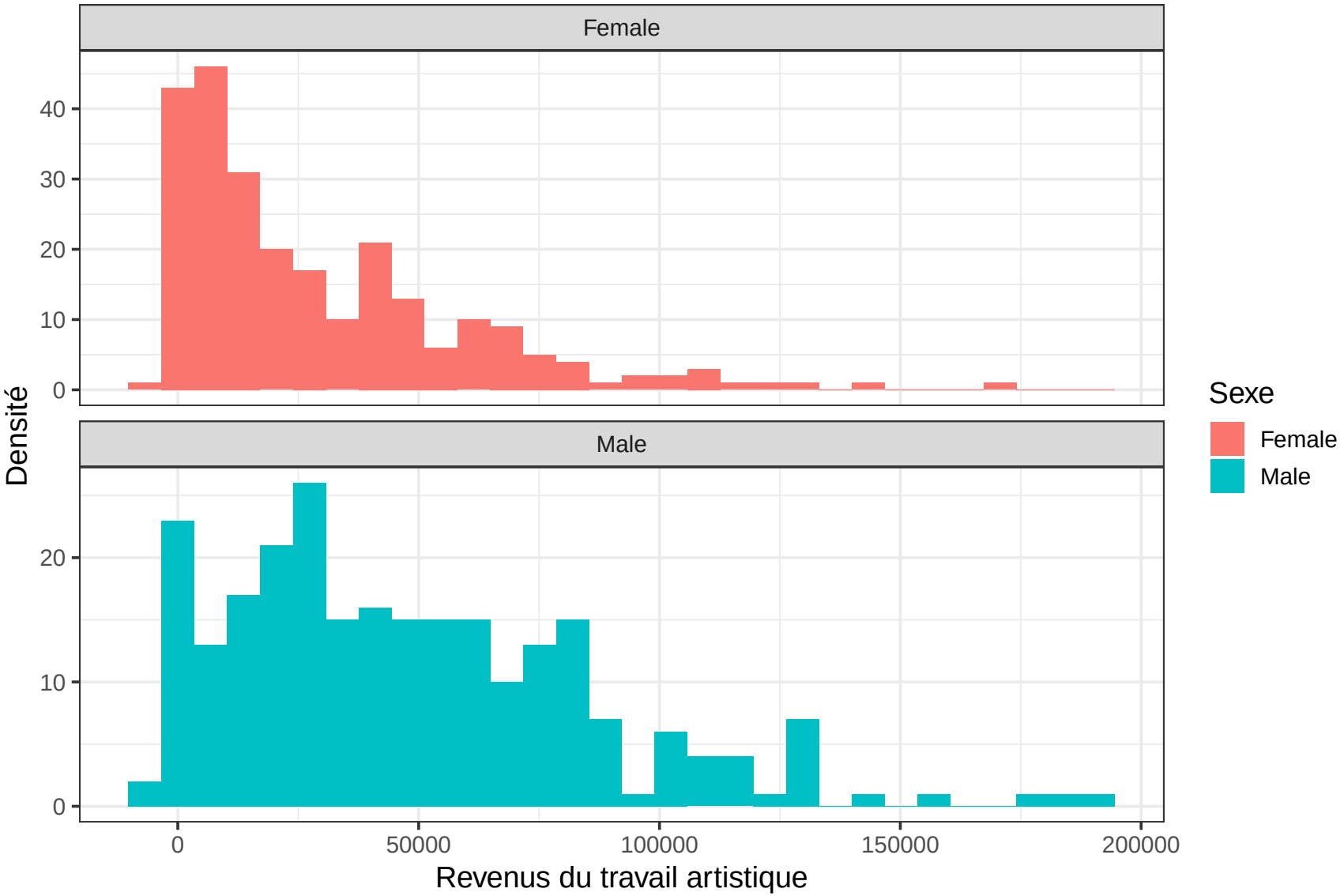
Comme pour la statistique univariée, on commence par regarder la distribution de la variable numérique dans les différentes catégories avant de calculer des indicateurs résumés par catégories et de les comparer entre eux.

On peut : comparer les distributions (histogrammes ou diagrammes de densité) ; comparer les indicateurs de tendances centrale et de dispersion (boites à moustaches) ; ou les deux (violins plots).

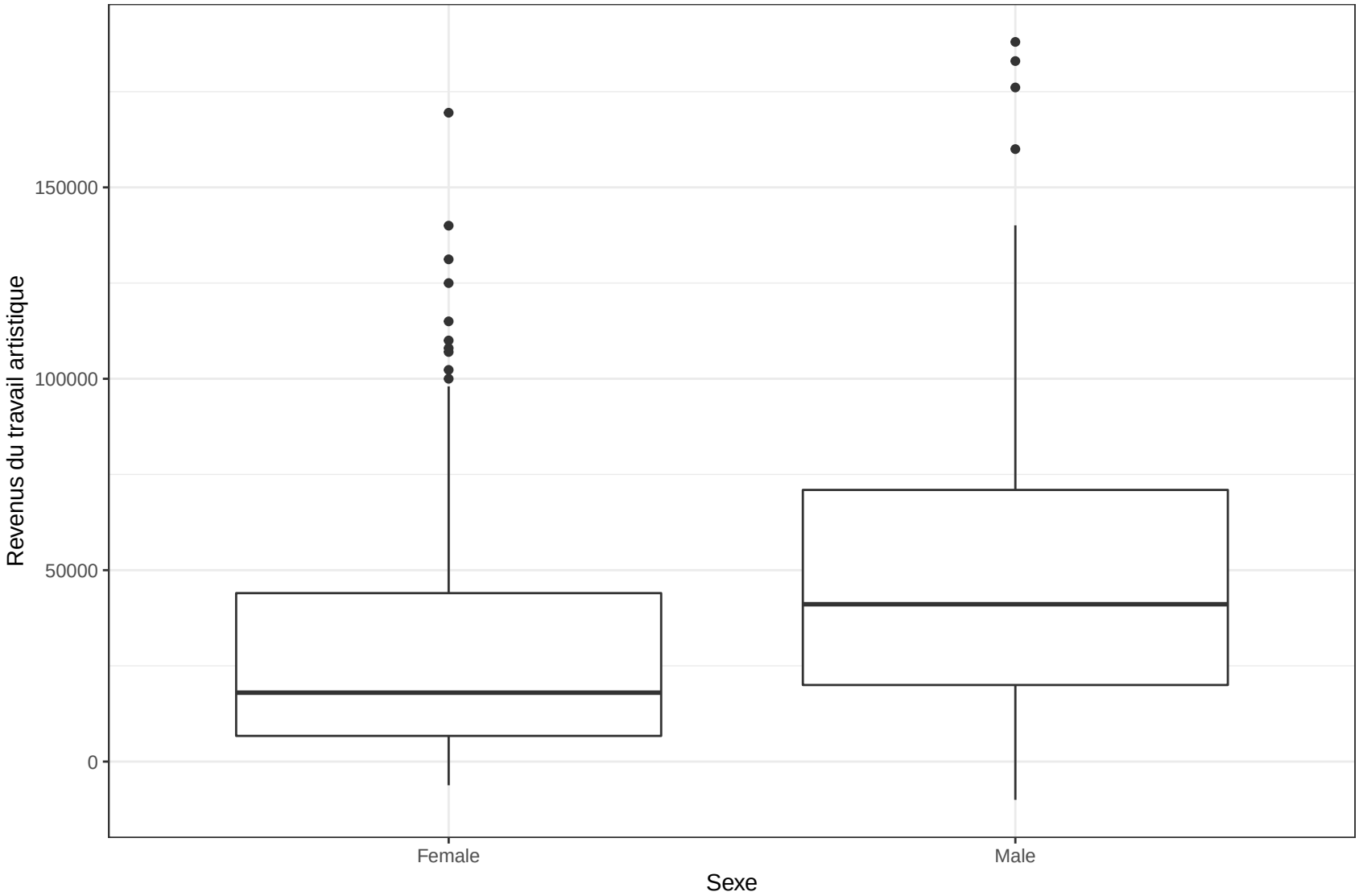
Revenus du travail artistique par sexe : densité



Revenus du travail artistique par sexe : histogramme



Revenus du travail artistique par sexe : boxplot



Revenus du travail artistique par sexe : violin plot

Sous-section 2

Comparaison

Comparaison

Une fois regardé la distribution, on peut déterminer quels sont les indicateurs de tendance centrale et de dispersion les plus pertinents, les calculer pour chaque groupe constitué par les observations appartenant à une modalité de la variable catégorielle, et les comparer entre eux.

Indicateur	Female	Male
Moyenne	33778	55751
Ecart_type	44219	66508
d1	1000	4500
q1	7500	18000
Mediane	22800	40000
q3	45400	70000
d9	74000	110000
rapport_interdecile	74	24