Séance 2 (1) : statistique descriptive univariée

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

Statistique descriptive

> *Ensemble des méthodes statistiques synthétisant les données afin de décrire une population.*

La statistique descriptive permet donc de passer des données **atomisées** (au niveau de l'individu) aux données **agrégées** (au niveau de la population).

Elle s'oppose à la statistique **inférentielle** qui recherche les liens entre les variables.

Statistiques descriptives

On peut décrire :

- une seule variable : statistique descriptive univariées (ce cours)
- deux variables : statistique descriptive bivariée (semaine prochaine)
- plusieurs variables : statistique descriptive multidimensionnelle

Les méthodes de la statistique descriptive comprennent :

- les tableaux
- les diagrammes
- les paramètres statistiques (valeurs numériques caractéristiques d'une variable quantitative)

Section 1

Description d'une variable catégorielle

Sous-section 1

Tri à plat

Tri à plat

Le **tri à plat** permet de décrire la **distribution** d'une variable catégorielle. Il s'agit de produire un tableau regroupant l'**effectif** et la **fréquence** de chacune des modalités d'une telle variable.

- **effectif** : nombre d'observations pour laquelle la variable étudiée prend la modalité en question. ($n_{modalite}$)
- **fréquence** : proportion de l'effectif d'une modalité par rapport à l'effectif total. ($\frac{n_{modalite}}{N}$)

Exemple de tri à plat

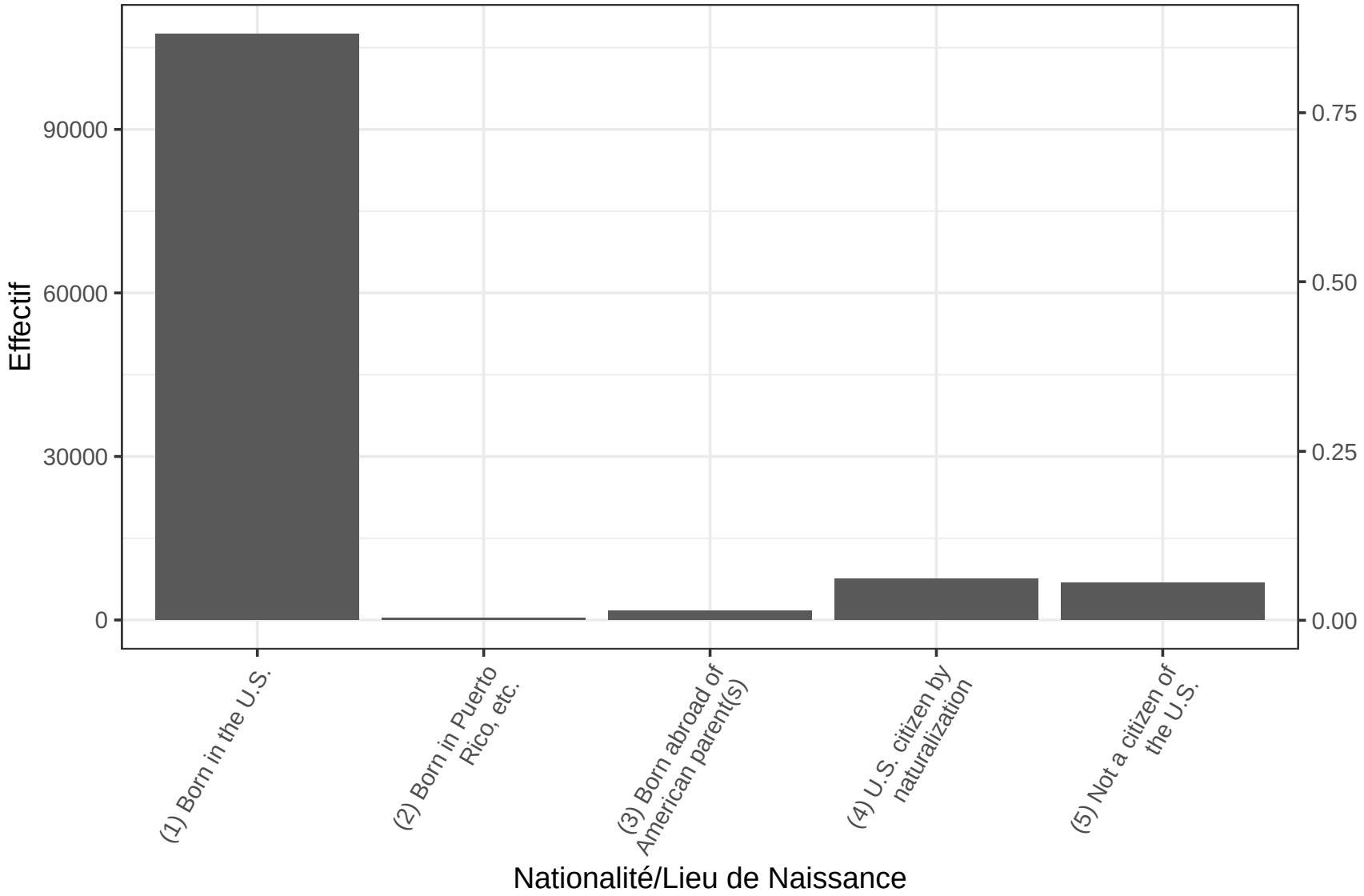
TAB. 1: Distribution de la nationalité/lieu de naissance des artistes américains

	Effectifs	Fréquence
(1) Born in the U.S.	107555	86.7
(2) Born in Puerto Rico, Guam, the U.S. Virgin Islands, or the Northern Marianas	356	0.3
(3) Born abroad of American parent(s)	1646	1.3
(4) U.S. citizen by naturalization	7607	6.1
(5) Not a citizen of the U.S.	6859	5.5
Total	124023	100

Représentation graphique

On représente graphiquement la distribution d'une variable catégorielle par un diagramme en barre ou barplot. Il peut se faire en effectifs ou en fréquence.

Représentation graphique



Mauvaises représentations graphiques

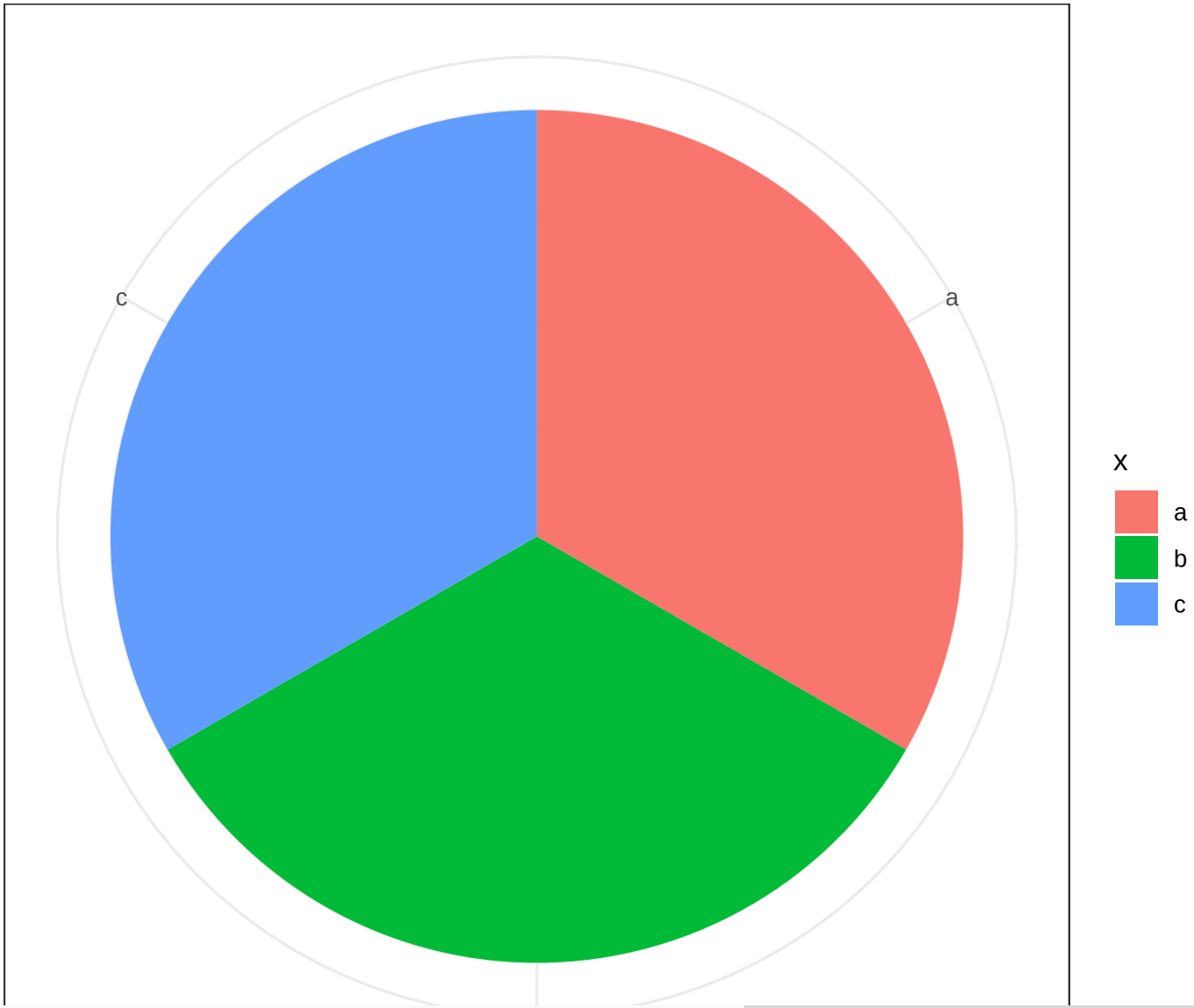
Une autre manière classique de représenter la distribution d'une variable catégorielle est le diagramme en "camembert" ou pie chart. Je vous conseille de l'éviter.

Un diagramme en barre représente les effectifs ou les fréquences par des *longueurs* ayant la *origine*. Elles sont faciles à comparer.

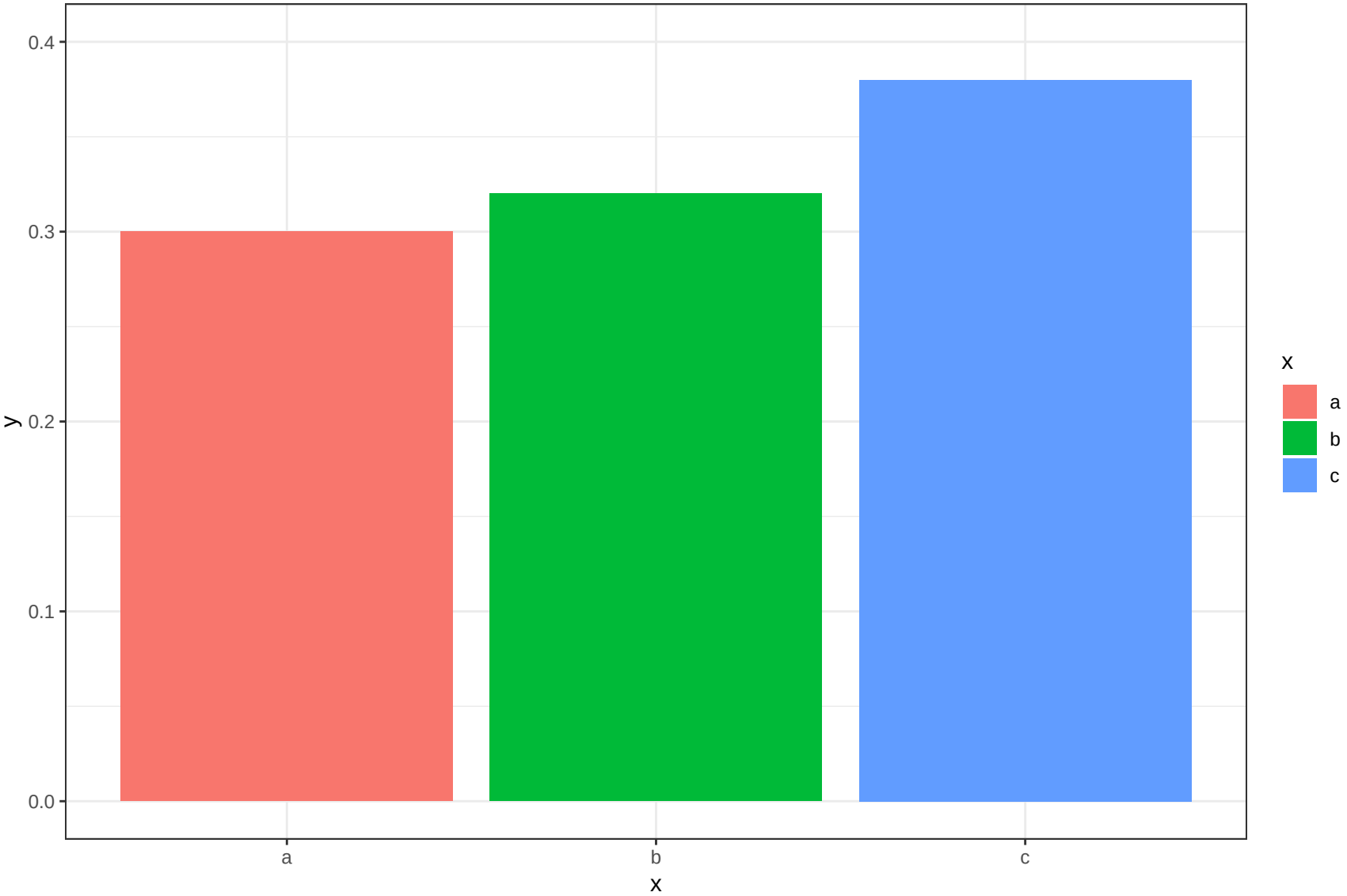
Un camembert les représente par des *angles* ayant des *origines* différentes. L'œil humain estime mal les angles et compare mal les mesures ayant des origines différentes.

Les logiciels comme excel proposent par défaut des options pour mettre ces deux types de diagrammes (en barre et en camembert) en 3D. La 3D rend plus difficile la comparaison des longueurs et encore plus trompeuse la comparaison des angles. Elle n'a aucune utilité.

Quelle est la catégorie la plus importante ?



Comparer à la slide précédente



Usage des couleurs dans les graphiques

Conventionnellement, on utilise :

- un jeu de couleur dégradé si on représente une variable catégorielle ordonnée. Le degradé de couleur renvoie à l'ordre des valeurs représentées. Cela vaut également pour une variable numérique, continue ou discrète.
- un jeu de couleur contrastée si on représente une variable catégorielle non ordonnée.

Attention à choisir les couleurs en gardant à l'esprit :

- les différences de perception (privilégier les palettes visibles par des personnes daltoniennes)
- les contraintes de publication (privilégier des couleurs se transformant en nuances de gris à l'impression noir et blanc).

Pour des conseils sur l'usage des couleurs et des suggestions de palettes adaptées aux différents usages : http://colorbrewer2.org

Les artistes américains parlent-ils anglais ?

Examinez le tableau suivant. Que peut-on dire de la maîtrise de la langue anglaise par les artistes américains ?

TAB. 2: How well does [the respondant] speaks english ?

	Effectifs	Fréquence
(1) Very well	12857	71.3
(2) Well	3401	18.9
(3) Not well	1453	8.1
(4) Not at all	319	1.8
Total	18030	100

Pouvez-vous repérer ce qui manque dans ce tableau ?

Les artistes américains parlent-ils anglais ?

TAB. 3: How well does [the respondant] speaks english ?

	Effectifs	Fréquence / NA	Fréquence sans NA
(1) Very well	12857	10.4	71.3
(2) Well	3401	2.7	18.9
(3) Not well	1453	1.2	8.1
(4) Not at all	319	0.3	1.8
NA	105993	85.5	NA
Total	124023	100	100

Examiner les valeurs manquantes

Les tris à plat peuvent être réalisés en incluant ou en excluant les valeurs manquantes de l'effectif total.

L'exclusion des valeurs manquantes est dangereuse car elle peut conduire à croire majoritaire une modalité qui ne l'est pas.

Par ailleurs, les valeurs manquantes ont des sens différents selon que

- il s'agit de l'effet d'un filtre lors du recueil des données : la variable sans objet pour l'observation (exemple : profession du conjoint d'une personne célibataire)
- il s'agit d'un manque d'information : la variable fait sens pour l'observation en question mais n'est pas renseignée (refus de répondre, absence de l'information dans les archives...)

On doit toujours commencer par étudier les valeurs manquantes ; on ne peut les exclure que si elles sont peu nombreuses et si elles sont liées à des filtres du questionnaire. Il faut alors mentionner la base du tableau (la composition de l'échantillon sur lequel porte le tri à plat).

Fréquences cumulées

Lorsque le tri à plat porte sur une variable catégorielle **ordonnée** (les modalités peuvent être triées dans un ordre logique), le tri à plat peut inclure, outre l'effectif et la fréquence, la **fréquence cumulée**.

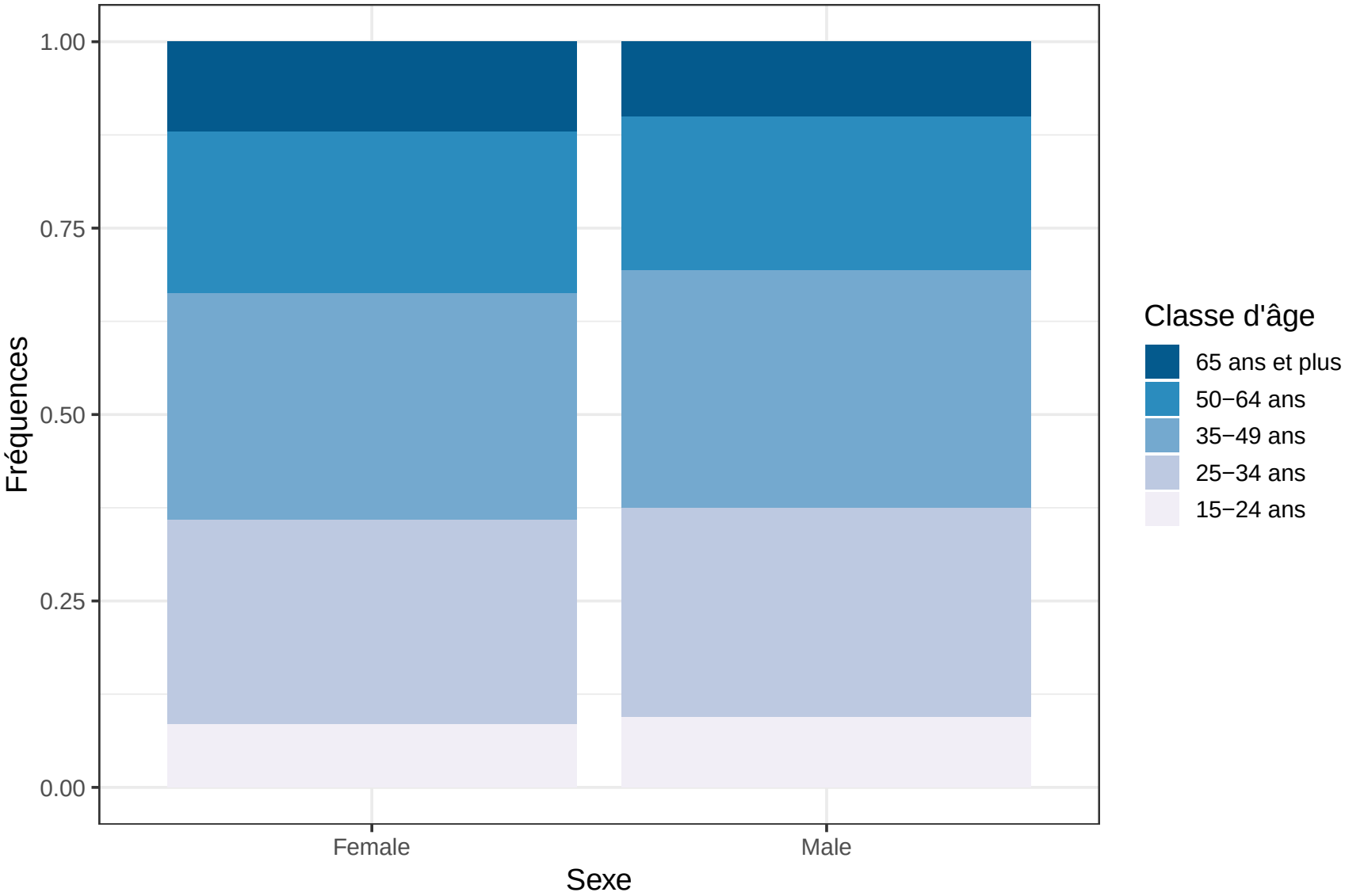
La fréquence cumulée d'une modalité est égale à la somme de sa fréquence et des fréquences de toutes les modalités précédentes dans l'ordre logique de classement des modalités.

Exemple de fréquence cumulée

TAB. 4: Distribution des âges des artistes américains

	n	%	%cum
15-24 ans	13656	11	11
25-34 ans	26181	21.1	32.1
35-49 ans	38632	31.1	63.3
50-64 ans	34389	27.7	91
65 ans et plus	11165	9	100

Représentation graphique d'une fréquence cumulée



Section 2

Description d'une variable numérique

Sous-section 1

Représentations graphiques

Décrire des variables quantitatives

La première chose à faire pour décrire des variables quantitatives est de les *visualiser* par des représentations graphiques. Elles permettent d'appéhender la forme générale de la distribution, d'estimer si celle-ci ressemble à une loi de distribution statistique connue, et de déterminer les meilleurs indicateurs synthétiques pour la résumer.

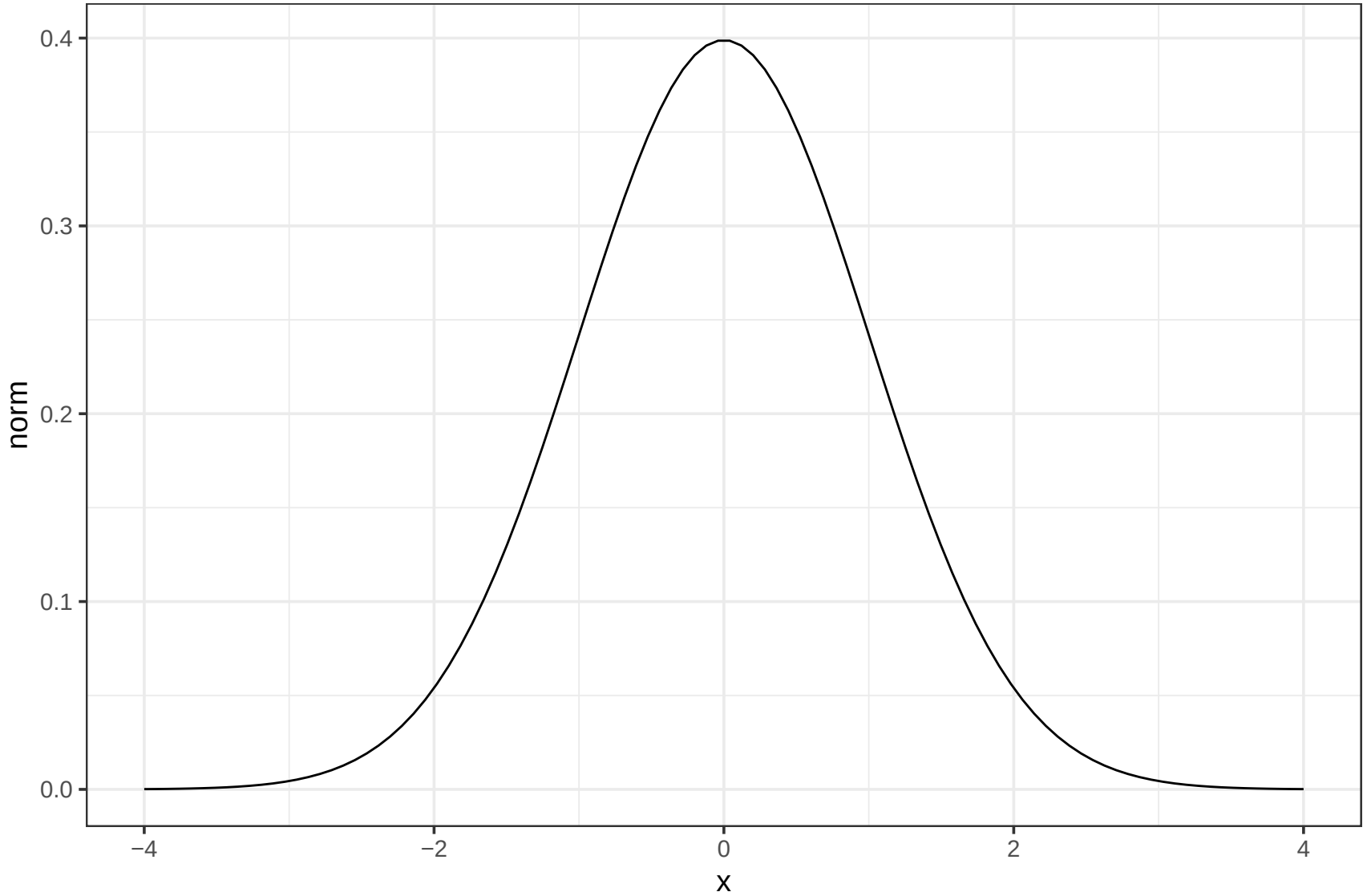
Représentations graphiques de variables quantitatives

Diagramme de densité (density plot). La densité est une fonction qui décrit la probabilité qu'une observation soit comprise dans un intervalle donné. La densité est le principal indice graphique pour reconnaitre une loi de probabilité.

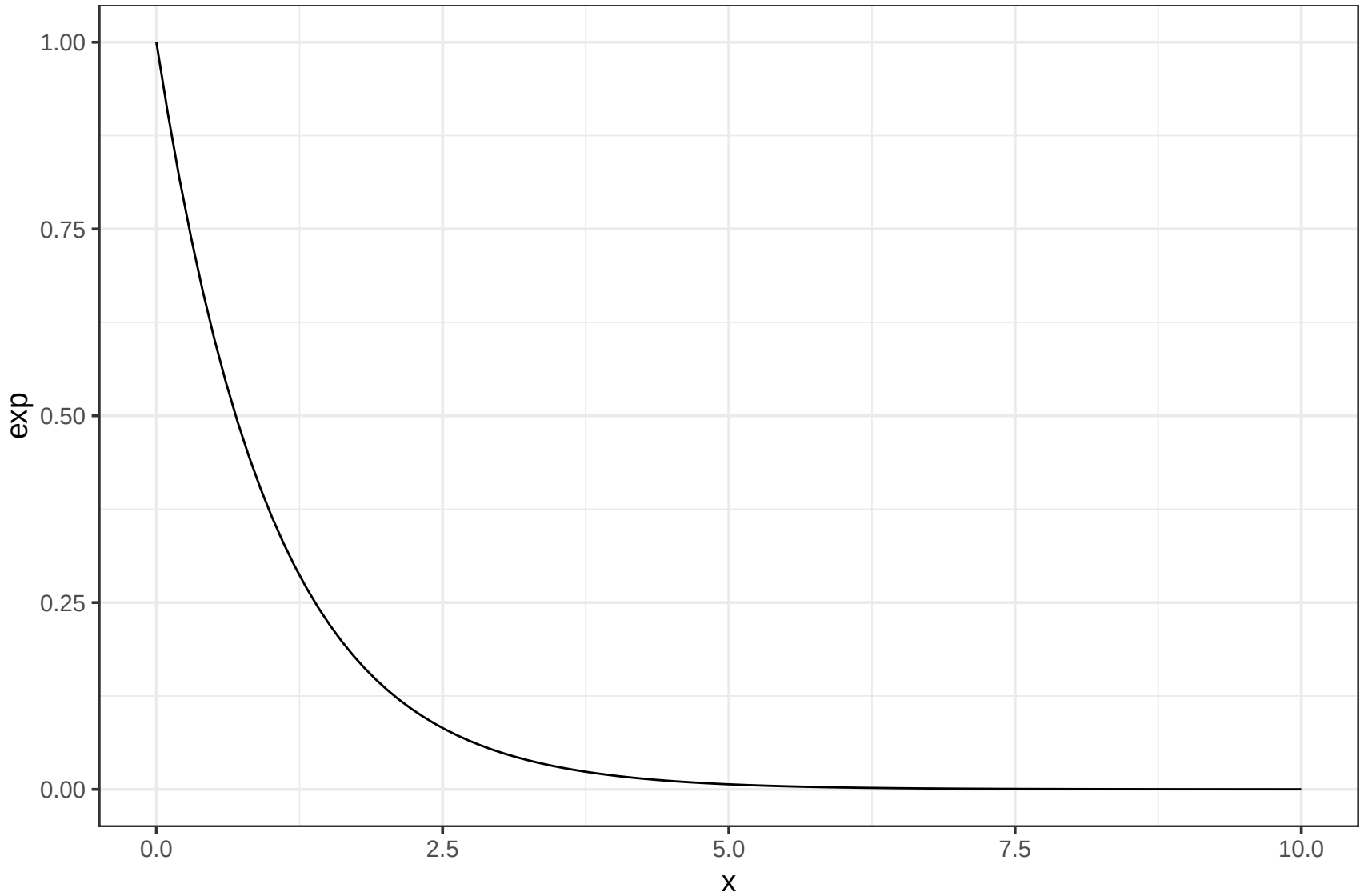
Histogramme (histogram). Découpage d'une variable numérique continue ou discrète en classes *d'amplitude égale*.

Boite à moustache (boxplot), **violin plot**: représentation de la distribution d'une variable numérique assortie de différents indicateurs de centralité et de dispersion.

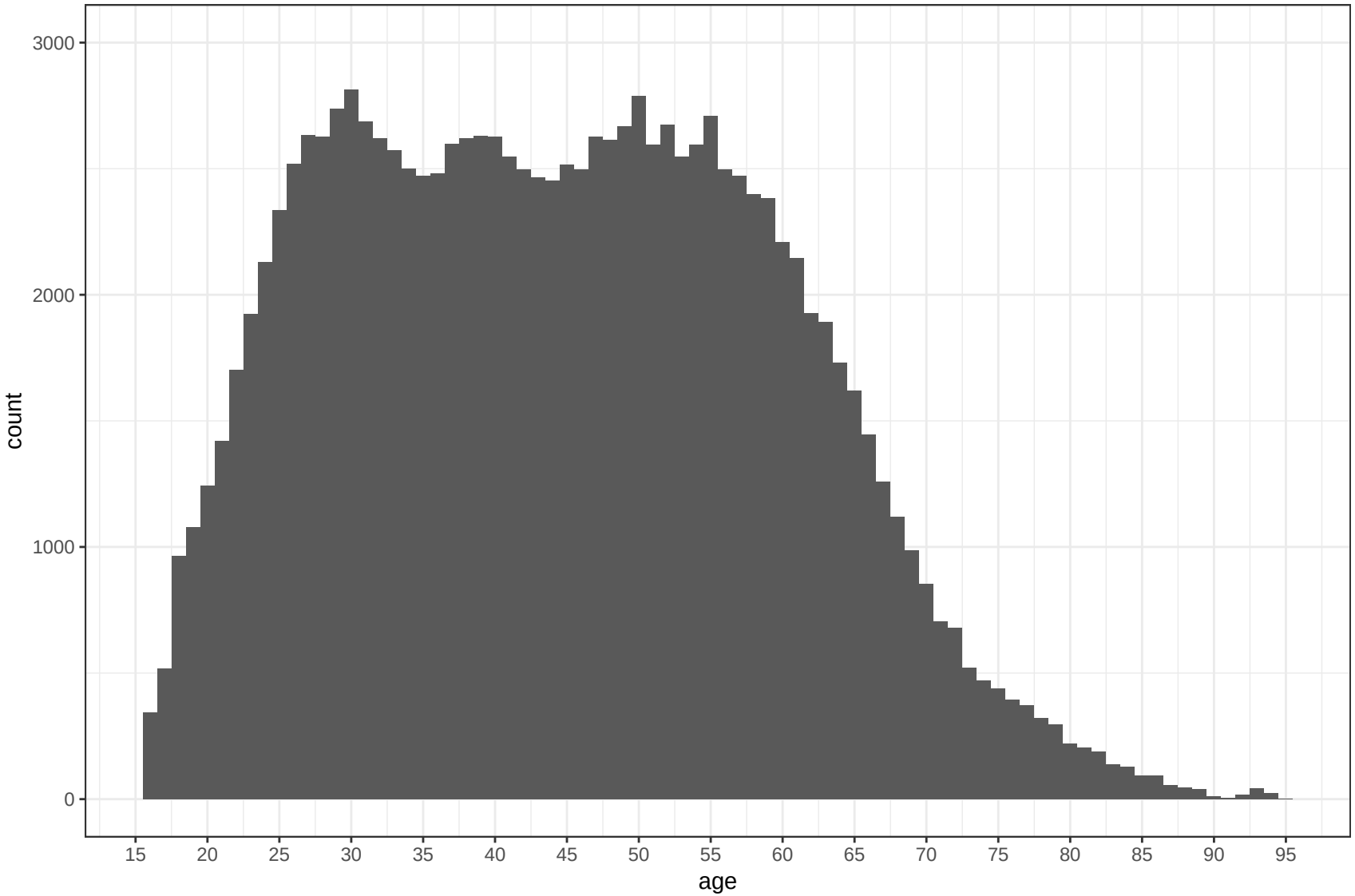
Densité d'une distribution normale



Densité d'une distribution exponentielle



Histogramme



Sous-section 2

Indicateurs de tendance centrale

Indicateurs de tendance centrale

Une fois les premiers diagnostics effectués, on peut calculer deux types d'indicateurs permettant de résumer les variables quantitatives :

- Les **indicateurs de tendance centrale** décrivent le cœur de la distribution. Ils désignent les valeurs autour desquels est concentrée la variable. Il en existe trois, le **mode**, la **moyenne** et la **médiane**.
- Les **indicateurs de dispersion** décrivent l'étendue de la distribution : la tendance des valeurs à s'étaler autour d'une valeur centrale. On trouve parmi eux la **variance** et l'**écart-type**.

Ces deux ensembles d'indicateurs sont complémentaires. Ils doivent être employés ensemble.

Le mode

Le mode est la modalité ou la valeur **la plus fréquente** d'une distribution.

Dans le cas d'une **variable qualitative**, le mode est la modalité qui a **l'effectif le plus élevé**.

Pour une variable quantitative discrète, le mode est la valeur qui a l'effectif le plus élevé.

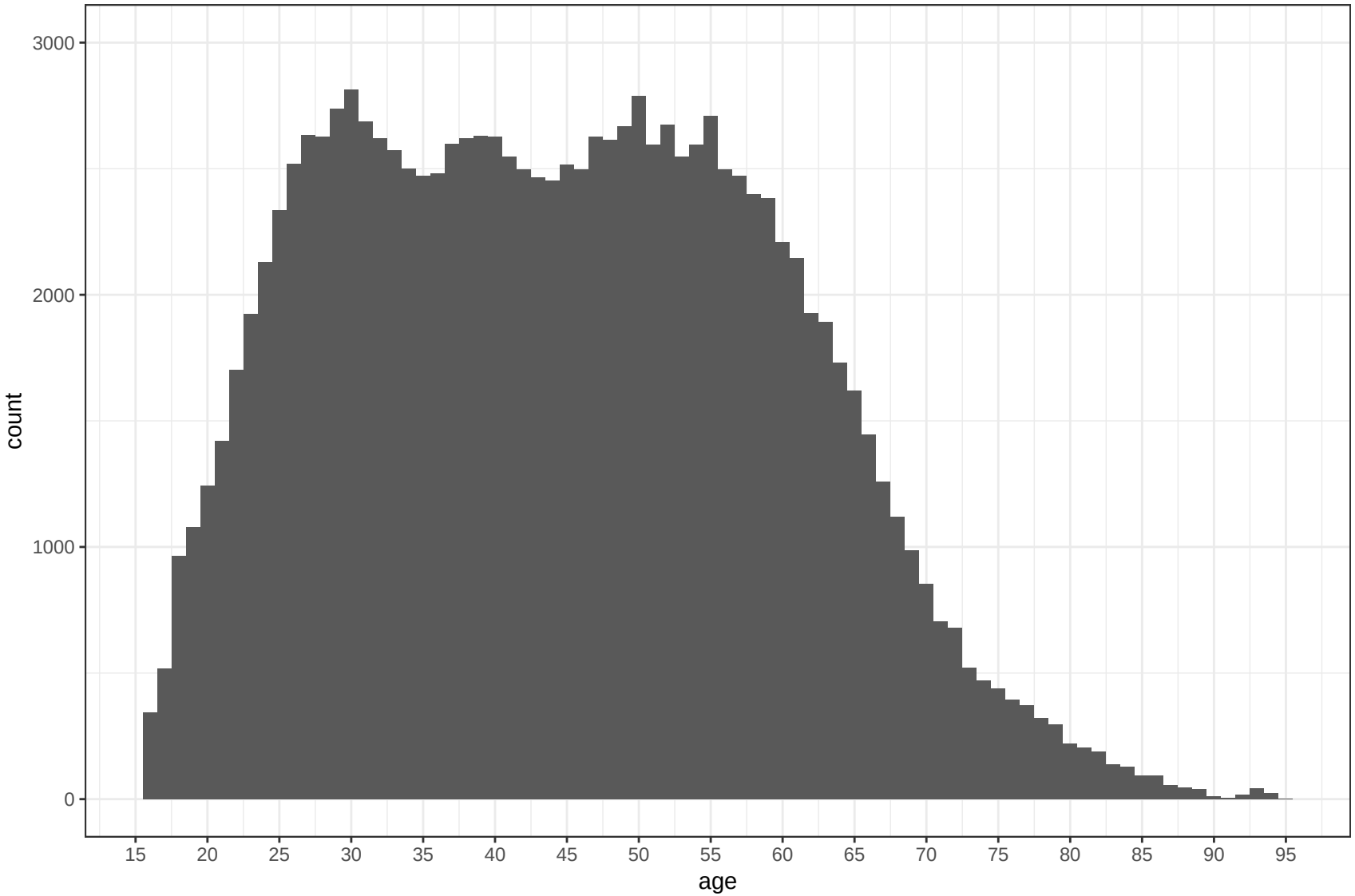
Pour une variable quantitative continue, le mode n'a aucun sens, car chaque valeur a une grande probabilité d'être unique. Dans ce cas, on ne peut le calculer qu'à condition de découper la variable en classes.

Mode d'une variable qualitative

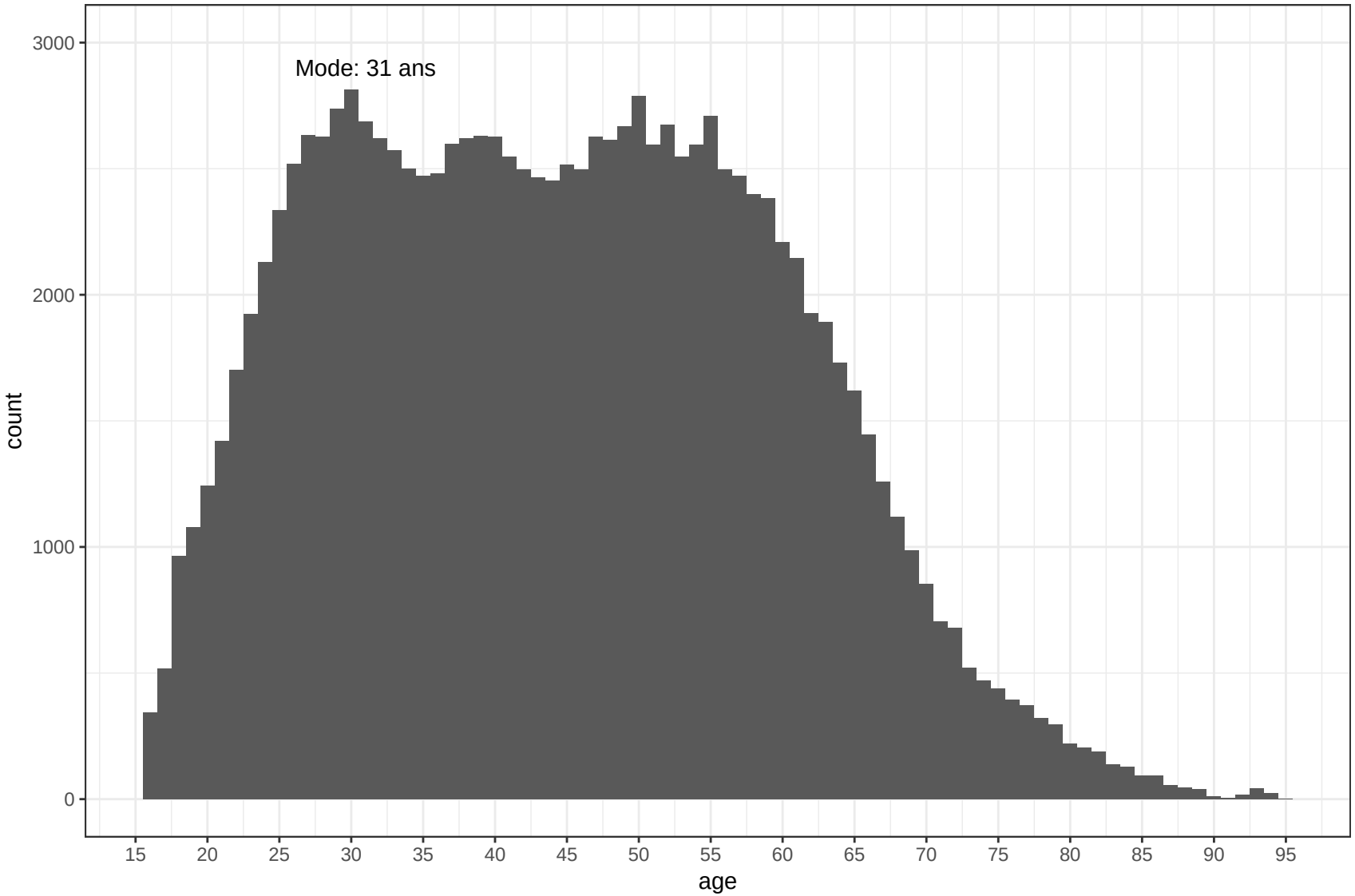
TAB. 5: Distribution de la nationalité/lieu de naissance des artistes américains

	Effectifs	Fréquence
(1) Born in the U.S.	107555	86.7
(2) Born in Puerto Rico, Guam, the U.S. Virgin Islands, or the Northern Marianas	356	0.3
(3) Born abroad of American parent(s)	1646	1.3
(4) U.S. citizen by naturalization	7607	6.1
(5) Not a citizen of the U.S.	6859	5.5
Total	124023	100

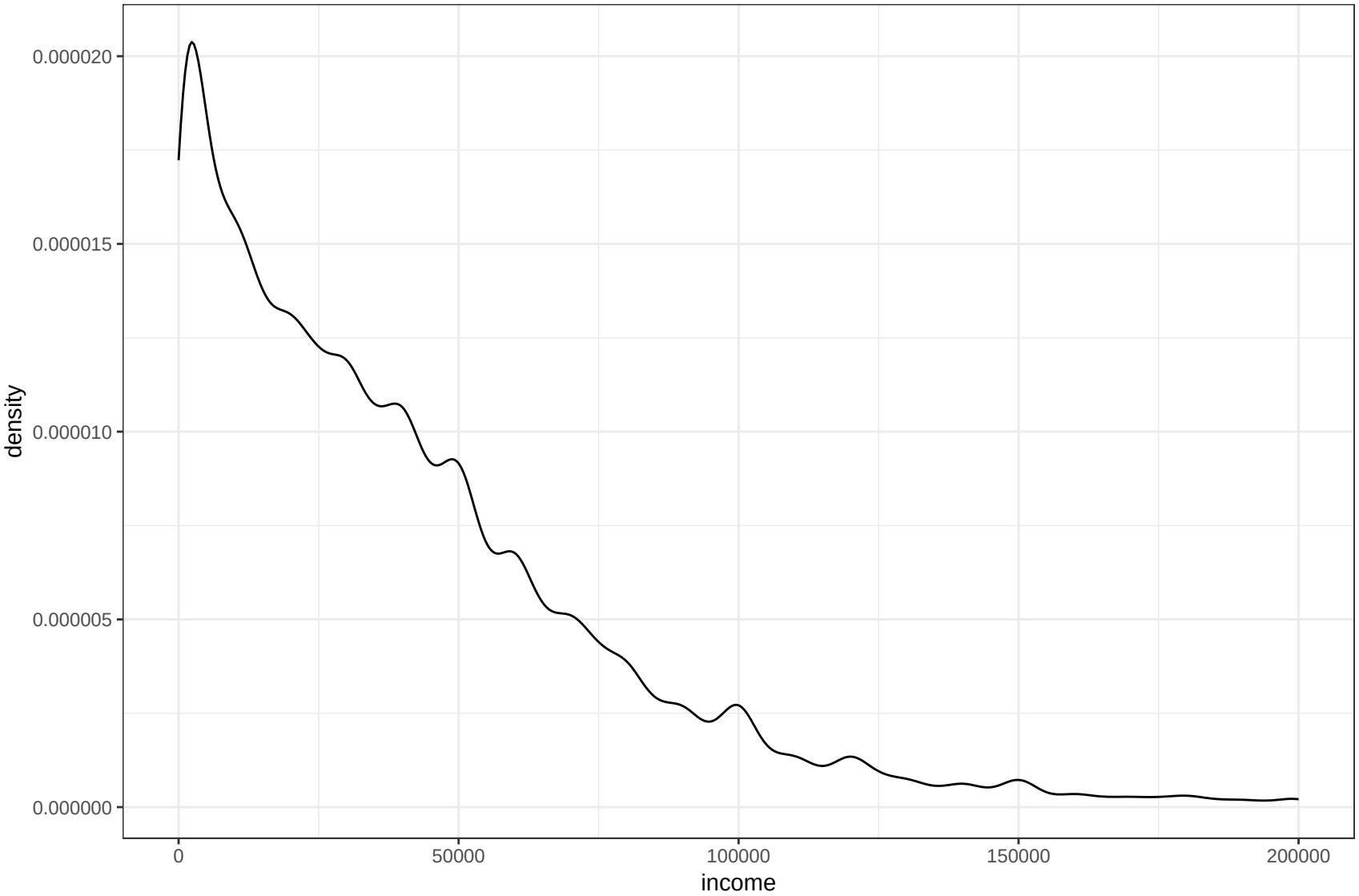
Mode d'une variable quantitative discrète



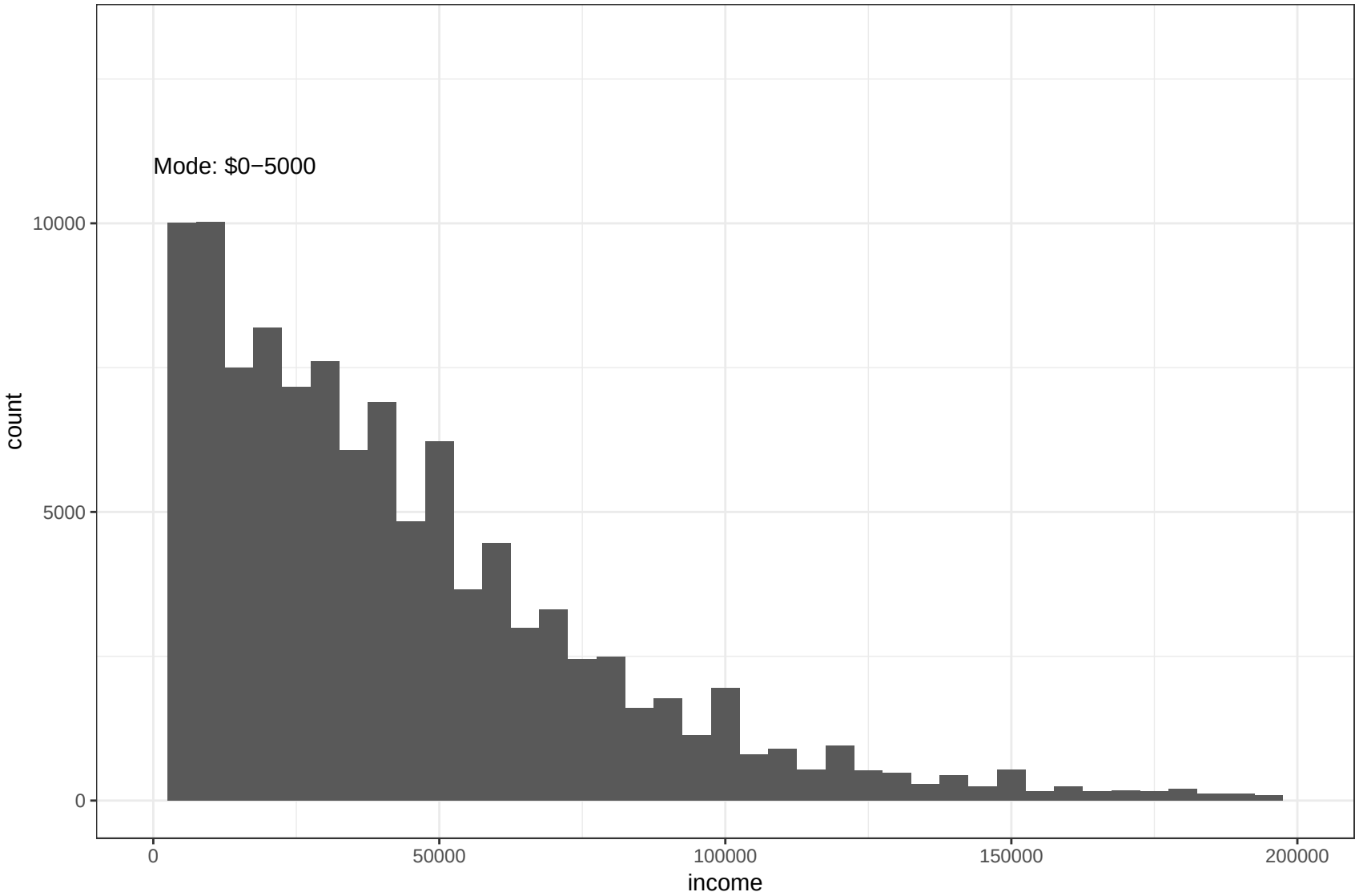
Mode d'une variable quantitative discrète



Mode d'une variable quantitative continue



Mode d'une variable quantitative continue



Médiane

La **médiane** est la valeur d'une variable telle que la moitié des valeurs lui est supérieure et l'autre moitié inférieure. Elle peut être calculée pour des **variables qualitatives ordonnée** ou pour des **variables quantitatives**.

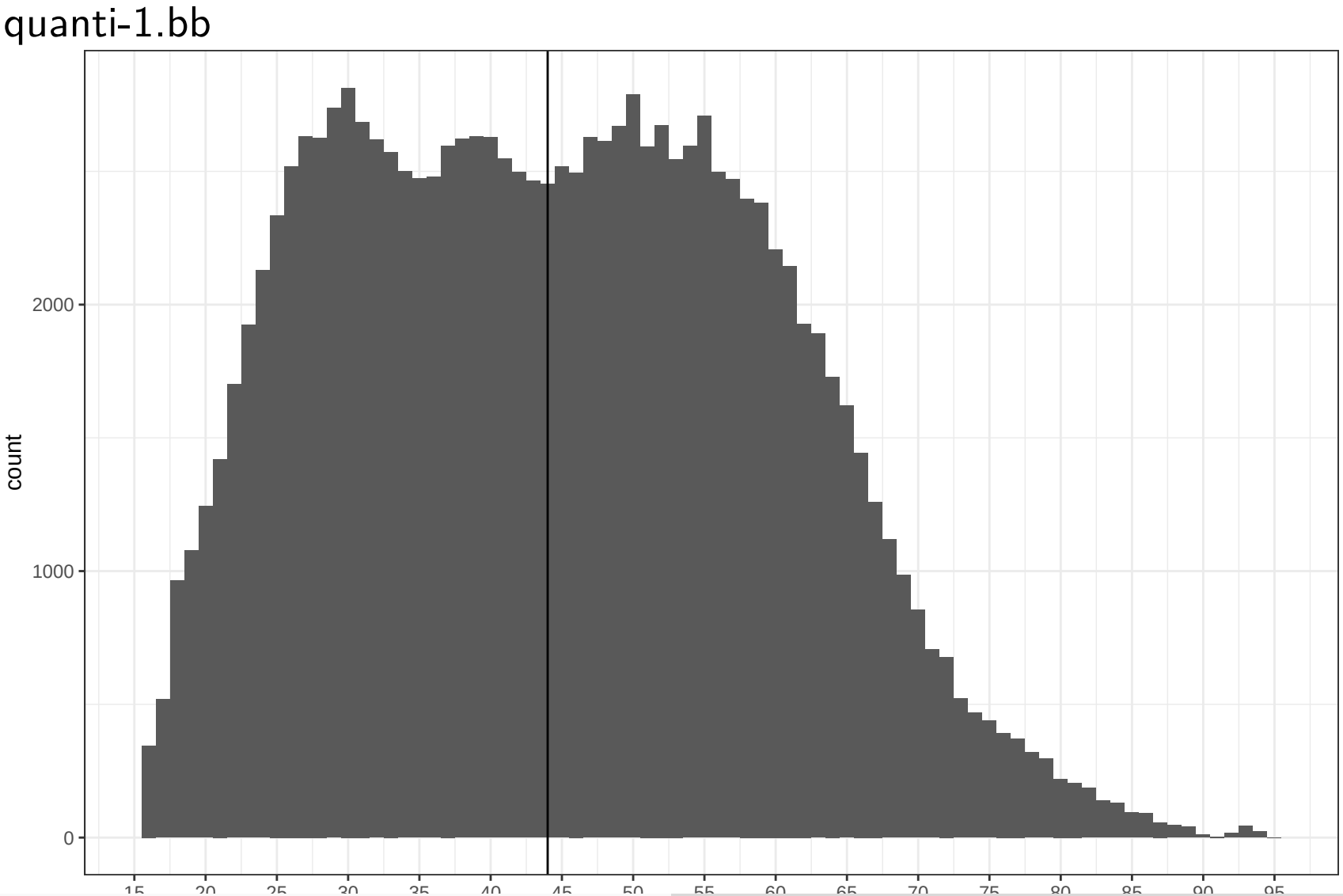
Médiane d'une variable ordonnée

Il s'agit de la classe qui contient la valeur séparant les observations ordonnées en deux sous-ensembles de taille égale. Ici, la médiane est la classe des 35-49 ans.

TAB. 6: Distribution des âges des artistes américains

	n	%	%cum
15-24 ans	13656	11	11
25-34 ans	26181	21.1	32.1
35-49 ans	38632	31.1	63.3
50-64 ans	34389	27.7	91
65 ans et plus	11165	9	100

Médiane d'une variable quantitative



Calcul d'une médiane

Il suffit de classer les observations par ordre croissant. Si la variable a un nombre d'observations n impair, la médiane est la valeur de rang $\frac{n+1}{2}$.

Si la variable a un nombre d'observations n pair, la médiane est la moyenne des observation de rang $\frac{n}{2}$ et $\frac{n}{2}+1$

Médiane d'une variable quantitative au nombre d'observation impair

Logement	Superficie
1	32
2	24
3	150
4	9
5	90
6	78
7	35
8	50
9	19

Médiane d'une variable quantitative au nombre d'observation impair

Logement	Superficie
4	9
9	19
2	24
1	32
7	35
8	50
6	78
5	90
3	150

La médiane est la valeur de rang $\frac{9+1}{2} = 5$, soit 35 m^2.

Médiane d'une variable quantitative au nombre d'observation pair

Logement	Superficie
1	32
2	24
3	150
4	9
5	90
6	78
7	35
8	50

Médiane d'une variable quantitative au nombre d'observation impair

Logement	Superficie
4	9
2	24
1	32
7	35
8	50
6	78
5	90
3	150

La médiane est la moyenne des valeurs de rang $\frac{8}{2} = 4$ et $\frac{8}{2} + 1 = 5$, soit $\frac{35+50}{2} = 42.5$ m^2.

La moyenne

La moyenne est la somme des valeurs d'une variable divisée par le nombre d'observations. On note $\bar{x}$ la moyenne de la variable x comprenant n observations X:

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} X_i$$

Elle ne s'applique qu'aux variables quantitatives.

Moyenne d'une variable quantitative

Logement	Superficie
4	9
2	24
1	32
7	35
8	50
6	78
5	90
3	150

La moyenne vaut ici $\frac{9+24+32+35+50+78+90+150}{8} = 58.5$ m^2.

Propriétés de la moyenne

La somme des écarts des valeurs à la moyenne est nulle. Soit $\bar{x}$ la moyenne d'une variable x de valeur X.

$$\sum_{i=1}^{n}(X_i - \bar{x}) = 0$$

La moyenne est très sensible à la présence de valeurs extrêmes.

Sous-section 3

Indicateurs de dispersion

Dispersion statistique

La dispersion statistique d'une distribution correspond à la tendance qu'ont les valeurs d'une variable à s'étaler autour d'une valeur centrale (moyenne, médiane, etc.). Cette tendance est inégale entre les variables.

Étendue

L'**étendue** d'une variable quantitative est la différence entre sa valeur maximum et sa valeur minimale.

$$\text{Étendue de X} = X_{max} - X_{min}$$

Quantile

Les **quantiles** sont les valeurs qui séparent les observations ordonnées d'une variable en x sous-ensembles de même effectif.

- La **médiane** est un quantile particulier qui sépare une distribution en deux sous-ensemble.
- Les **quartiles** séparent la distribution en quatre sous-ensemble.
- les **déciles** en dix sous-ensembles
- les **centiles** en cent sous-ensembles...

Quartiles

Les **quartiles** sont donc les trois valeurs qui découpent une distribution en quatre classes d'effectifs égaux.

25%	50%	75%
32	44	56

Déciles

Les **déciles** sont les neuf valeurs qui découpent une distribution en dix classes d'effectifs égaux.

10%	20%	30%	40%	50%	60%	70%	80%	90%
25	30	34	39	44	49	54	59	65

Intervalle et rapport

Les quantiles permettent de calculer deux types d'indicateurs de dispersion :

- l'**intervalle interquantile** désigne la différence entre la valeur du dernier quantile et la valeur du premier quantile.
- le **rapport interquantile** désigne le rapport de la valeur du dernier quantile sur la valeur du premier quantile.

Les inégalités de revenu entre les artistes américains

25%	50%	75%
11020	30270	59400

L'**intervalle interquartile** vaut 59400 - 11020 = 48380 dollars. Les 25% d'artistes les mieux payés touchent au moins 48380 dollars **de plus** que les 25% d'artistes les moins biens payés.

Les inégalités de revenu entre les artistes américains

10%	20%	30%	40%	50%	60%	70%	80%	90%
2000	8000	15000	22600	30270	40000	50600	67000	95000

Le **rapport interdécile** vaut $\frac{95000}{2000} = 47.5$. Les 10% d'artistes les mieux payés touchent 47.5 **fois plus** que les 10% d'artistes les moins bien payés.

Variance

La **variance** est égale à **la moyenne des carrés des écarts à la moyenne**. On la note σ^2. Soit $\underline{\sigma}_x^2$ la variance de la variable x, comprenant n observations, et $\overline{(x)}$ sa moyenne.

$$\sigma_x^2 = \frac{1}{n} * \sum_{i=1}^{n} (X_i - \bar{x})^2$$

L'écart-type

L'**écart-type** est **la racine carrée de la moyenne des carrés des écarts à la moyenne**, c'est-à-dire **la racine carrée de la variance**. On le note σ. Ainsi :

$$\sigma_x = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (X_i - \bar{x})^2} = \sqrt{\sigma_x^2}$$

Calcul de variance et d'écart type

Logement	Superficie	Valeur - moyenne	(Valeur - moyenne)^2
4	9	-49.5	2450
2	24	-34.5	1190
1	32	-26.5	702.2
7	35	-23.5	552.2
8	50	-8.5	72.25
6	78	19.5	380.2
5	90	31.5	992.2
3	150	91.5	8372

$$\sigma_x^2 = 1839$$

$$\sigma_x = \sqrt{\sigma_x^2} = 42.88$$

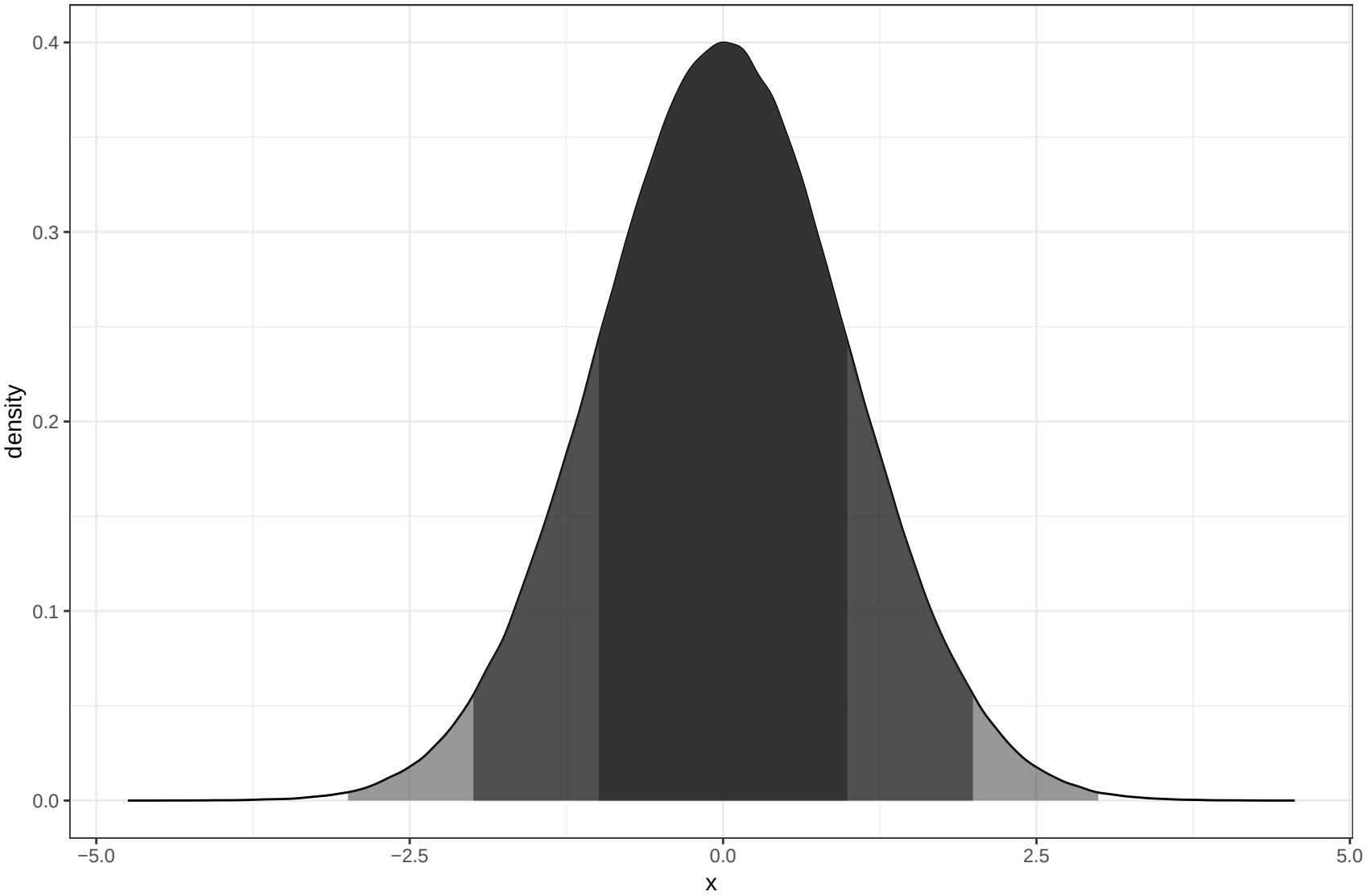
Interprétation

Si l'écart-type est sensiblement plus grand que la moyenne, la dispersion est importante.

Si la distribution est approximativement normale, alors :

- 68,3% des valeurs sont comprises dans $X \pm \sigma$
- 95,5% des valeurs sont comprises dans $X \pm 2\sigma$
- 99,7% des valeurs sont comprises dans $X \pm 3\sigma$

Loi normale



Sous-section 4

Représentations graphiques synthétiques

Comment représenter graphiquement des statistiques descriptives

Les bonnes représentation graphiques d'une variable numérique font apparaît à la fois la distribution de la variable et des indicateurs résumées. Quoi qu'il en soit, elles ne peuvent remplacer les diagrammes de densité et les histogrammes.

Les boites à moustaches (boxplot) et les violin plot remplissent ce rôle.

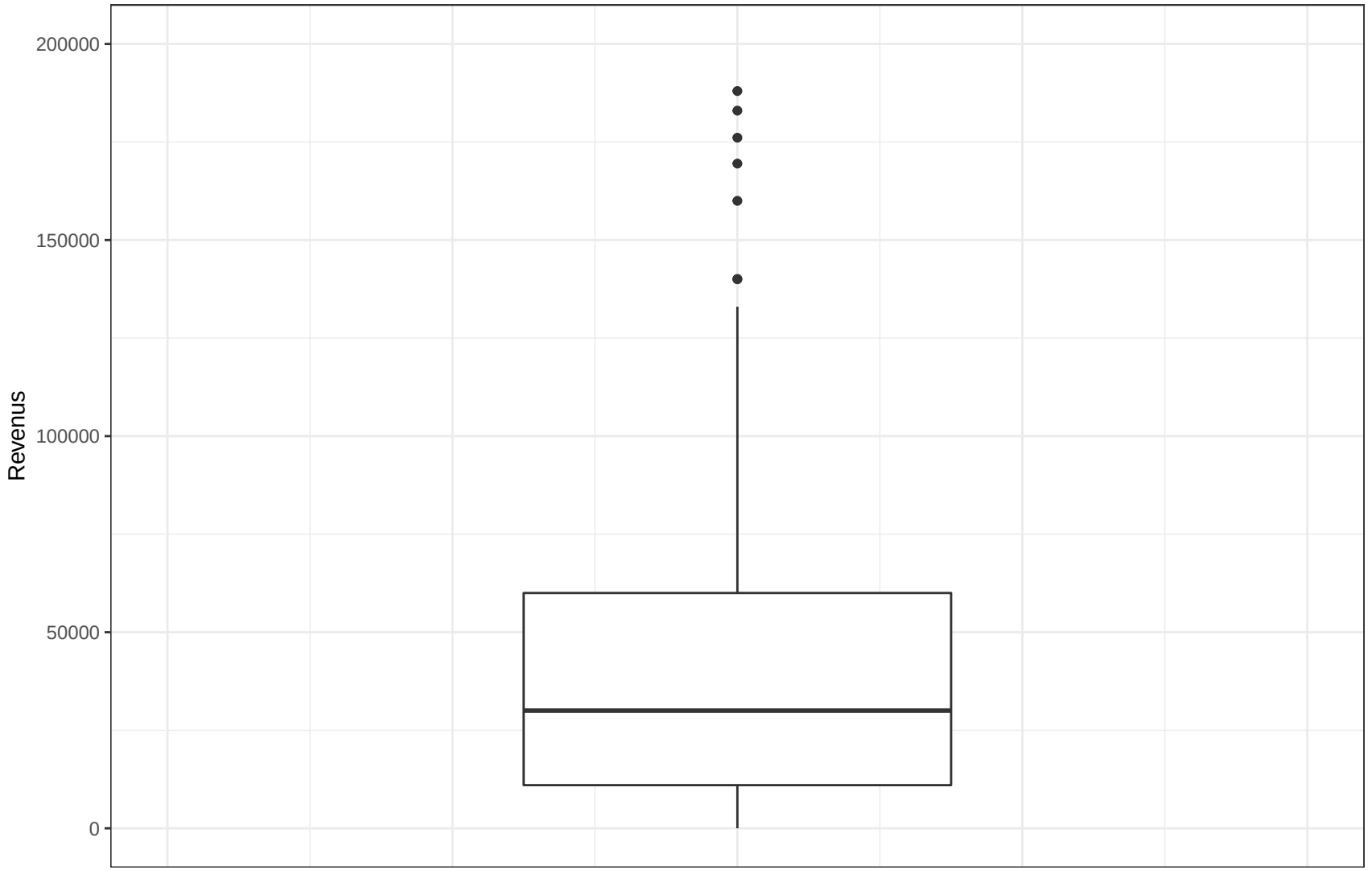
Une mauvaise manière de résumer l'information consiste à se contenter de représenter la tendance centrale. Par exemple, il est possible de représenter par un diagramme en barre la moyenne d'une variable numérique, mais on perd alors énormément d'information (en particulier la dispersion de cette variable).

Boite à moustache

La **boite à moustache** (*box and whisker plot* ou *boxplot*) est un diagramme présentant plusieurs indicateurs de tendance centrale et de dispersion d'une variable quantitative. On y trouve en particulier :

- la **médiane** est un trait gras au milieu de la boite
- les premiers et troisième **quartiles** délimitent une boite autour de la médiane
- les **moustaches** (*whiskers*) s'étendent des deux côtés de la boite jusqu'à $1,5 * intervalleinter - quartile$
- les **valeurs extrêmes** (*outliers*), les valeurs qui ne sont pas comprises à l'intérieur des moustaches, sont représentées par des points

Revenus des artistes



Revenus des artistes

Violin plot : Revenus des artistes