Analyse statistique - Séance 3 : La représentation graphique

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2023

La représentation graphique des données

La *représentation graphique des données* sert à communiquer de l'information par le biais visuel plutôt qu'écrit.

- Les graphiques *de présentation* sont définitifs, synthétiques et de haute qualité : ils doivent résumer de manière efficace et convaincante l'information.
- Les graphiques *exploratoires* sont temporaires, incomplets mais rapides à construire : ils permettent de prendre connaissance des données pour formuler ou vérifier des hypothèses.
- Ces deux types de représentations ne répondent donc pas aux mêmes exigences de lisibilité, d'annotation et de légendage.

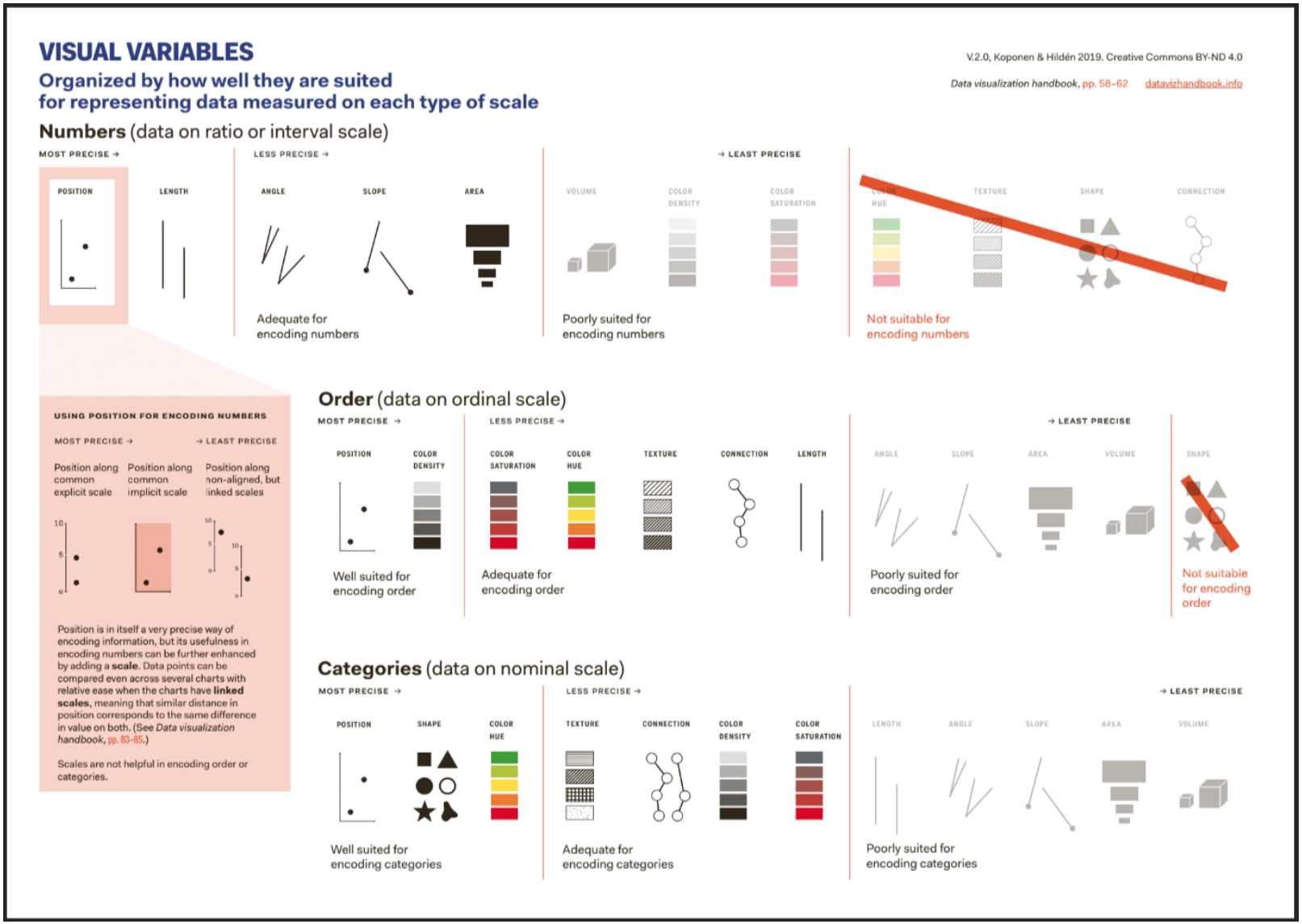
La représentation graphique des données

Le choix du type de réalisation graphique est lié à la nature de l'information représentée, limité ici à :

- La représentation de quantités et de distributions par des *histogrammes*, *diagrammes en barres*.
- La comparaison de quantités et de distributions par des *histogrammes*, *diagrammes en barres* et des *boîtes à moustache*.
- La corrélation par des *nuages de point* .
- L'évolution temporelle d'un indicateur par des *courbes*.

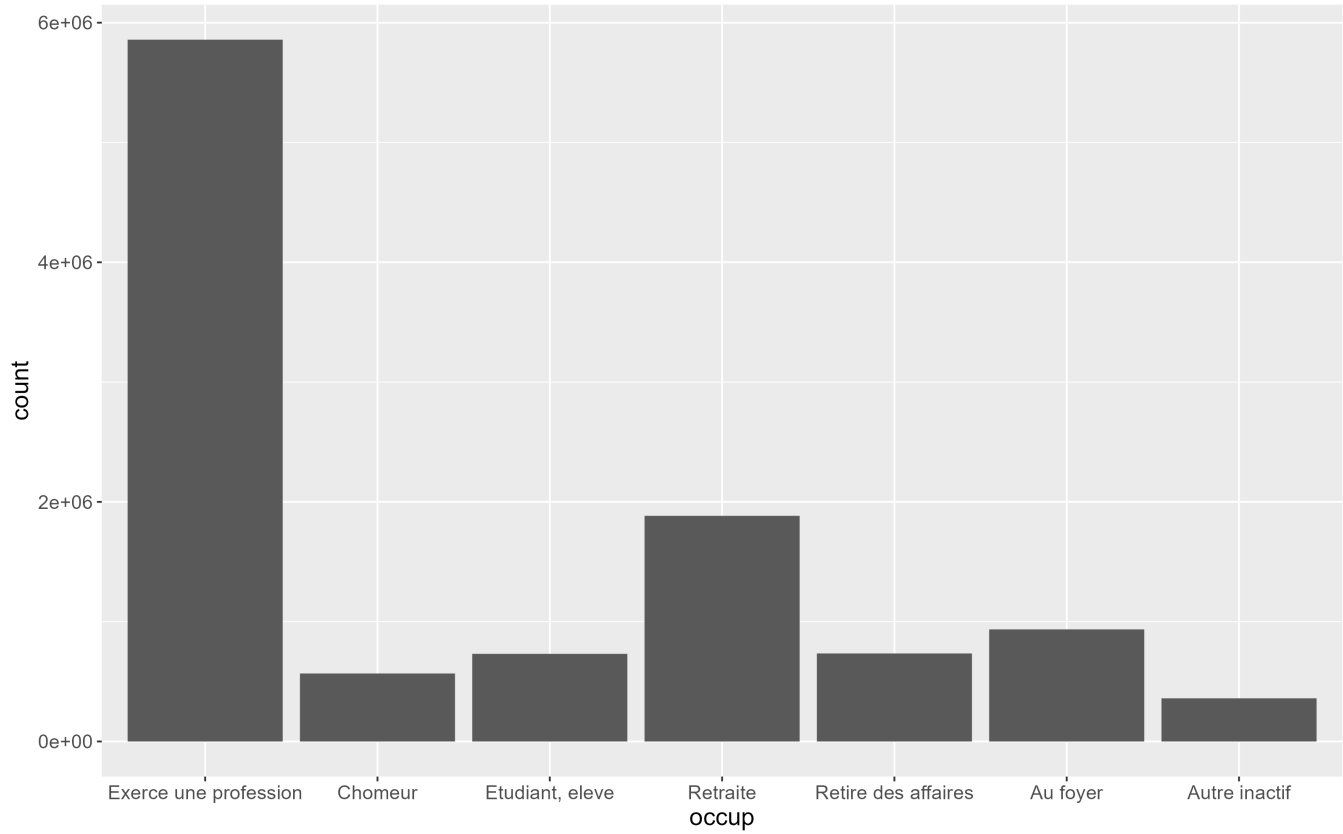
En parallèle, on présentera les règles de construction des tableaux d'effectifs et de proportions.

Quelques repères



Les diagrammes en barres

Un *diagrame en barres* permet de représenter la répartition des effectifs d'une variable catégorielle en construisant des barres d'une hauteur proportionnelle au nombre d'individus de chaque modalité.
Quelles informations et améliorations esthétiques manquent à ce diagramme ?



Représenter les effectifs d'une variable catégorielle

Le diagramme précédent est la traduction graphique d'un tri à plat qui présente les effectifs des modalités d'une variable catégorielle, la population étudiée, la source des données et qui s'accompagne d'un guide de lecture.

Répartition de l'occupation principale de la population en 2003

	Exerce une profession	Chomeur	Etudiant, eleve	Retraite	Retire des affaires	Au foyer	Autre inactif
Effectifs en milliers de personnes	5 858	568	731	1 884	735	936	360

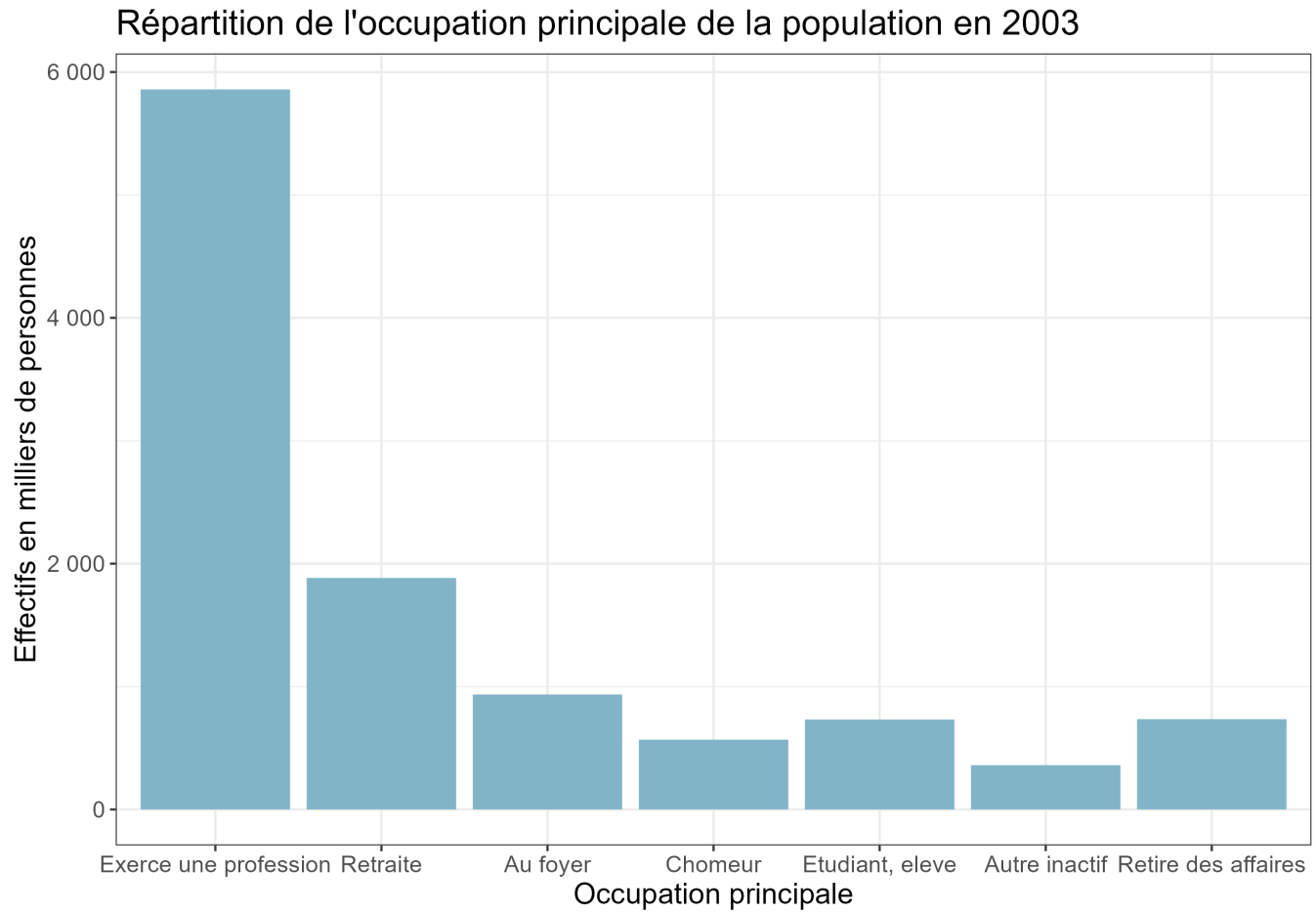
Lecture : en 2003, 5 858 milliers de personnes exercaient une profession et 1 884 milliers de personnes étaient à la retraite

Champ : Individus de 18 ans et plus habitant en France métropolitaine

Source : Histoire de Vie 2003 (sous-échantillon de 2000 individus) | A. POISSONNIER | 2024

Les diagrammes en barres

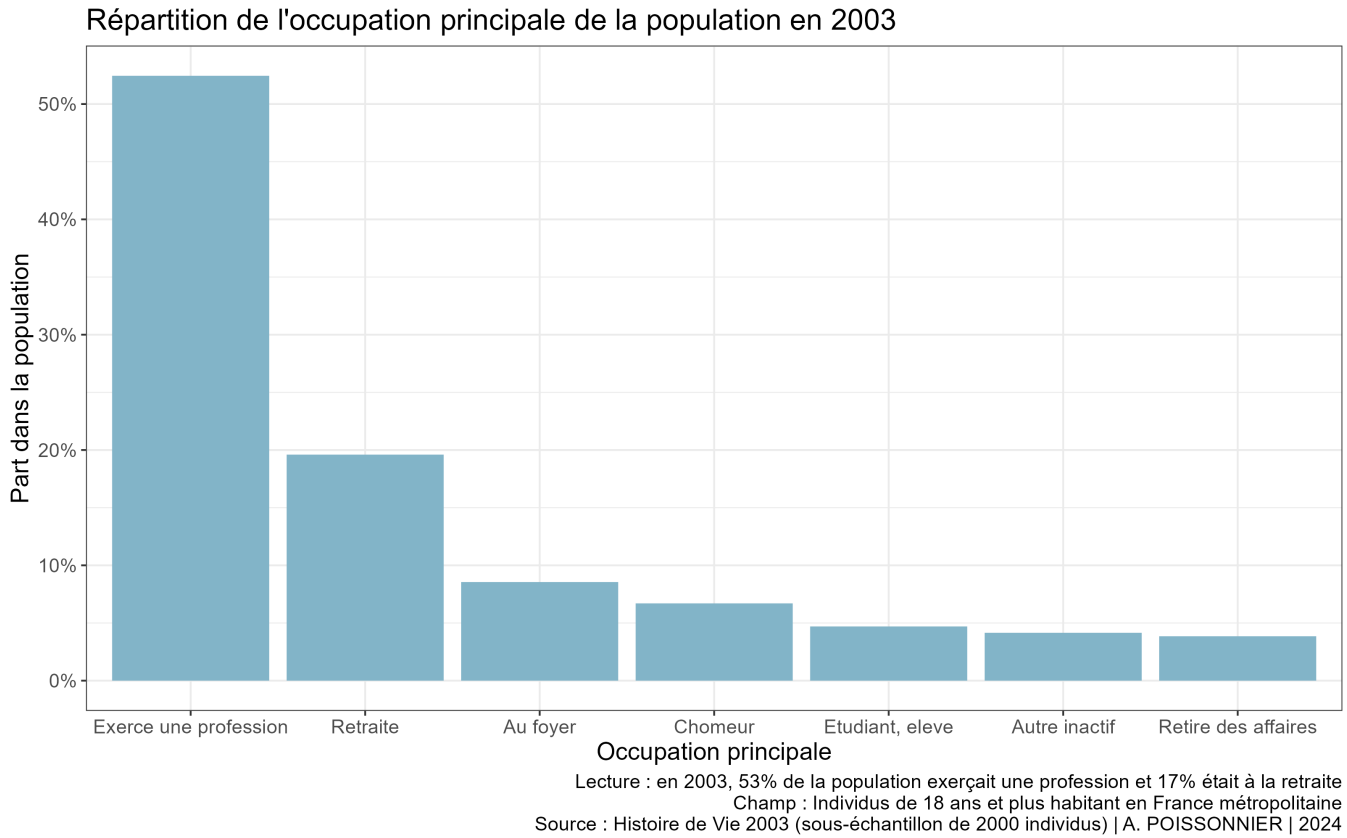
Le graphique doit être lisible et explicite, c'est-à-dire doté d'un titre, d'une légende et d'axes précisant les unités. L'ordre des modalités respecte ici la hiérarchie entre les effectifs.



_ecture : en 2003, environ 5,8 millions de personnes exercaient une profession et 1,8 millions de personnes étaient à la retraite
Champ : Individus de 18 ans et plus habitant en France métropolitaine
Source : Histoire de Vie 2003 (sous-échantillon de 2000 individus) | A. POISSONNIER | 2024

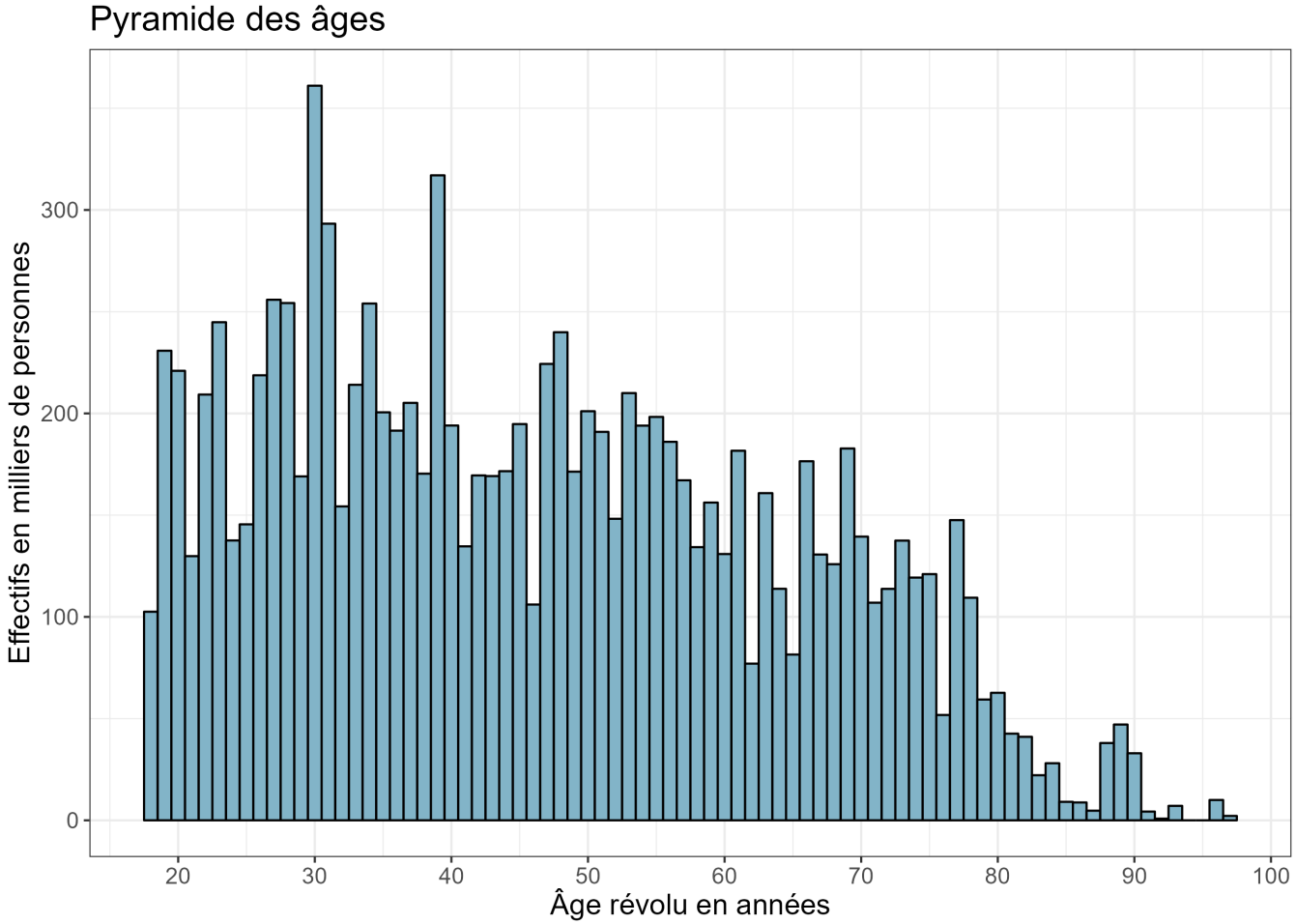
Les diagrammes en barres

Les effectifs bruts ne sont pas toujours nécessaires et le diagramme peut à la place représenter la fréquence ou la part de chaque modalité .



Les histogrammes

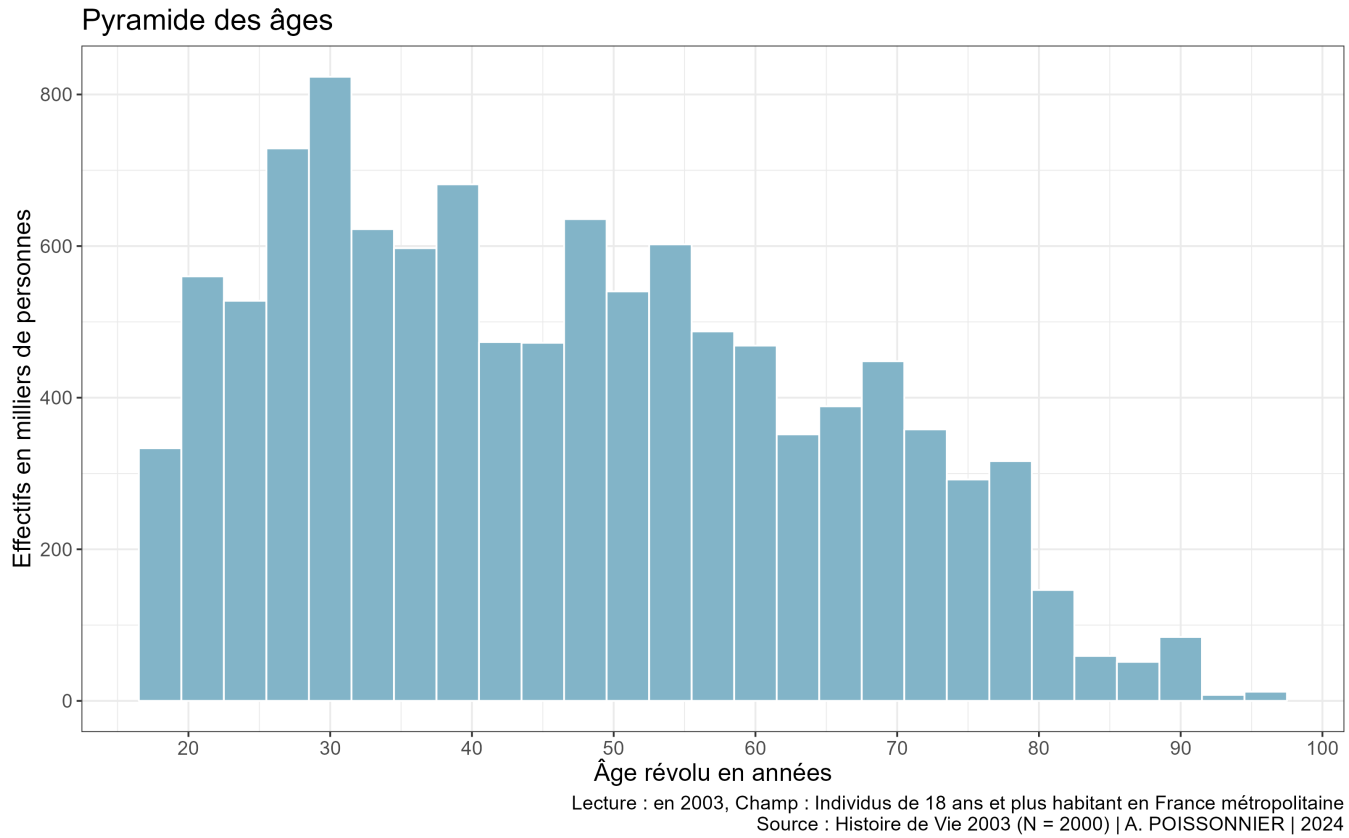
Un *histogramme* permet de représenter la distribution d'une variable numérique continue en construisant des barres d'une hauteur proportionnelle au nombre d'individus de chaque valeur prise par la variable.



Lecture : en 2003, Champ : Individus de 18 ans et plus habitant en France métropolitaine
Source : Histoire de Vie 2003 (N = 2000) | A. POISSONNIER | 2024

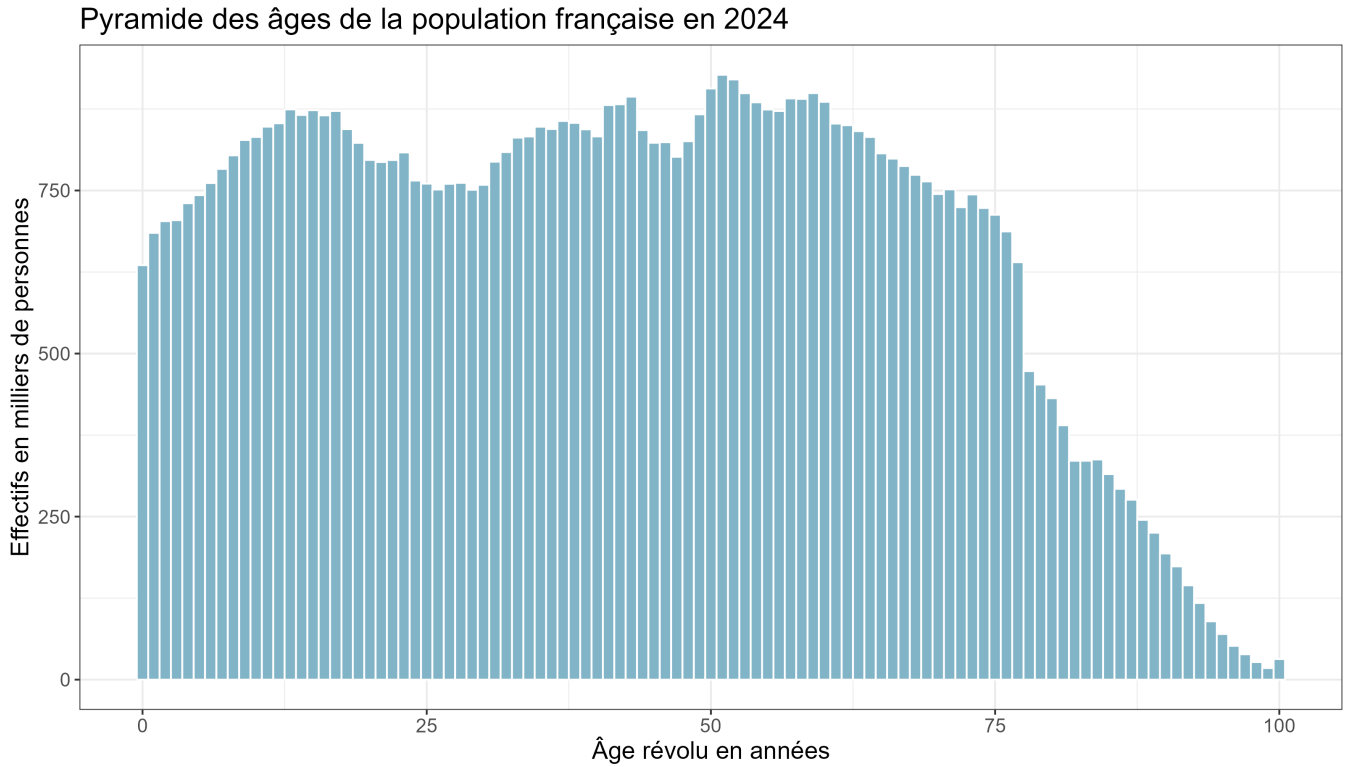
Les histogrammes

La précision de la mesure et la taille de l'échantillon déterminent le niveau de granularité de la représentation, ce qui impose parfois de regrouper plusieurs valeurs ensemble comme ici l'âge en ensembles de 3 années.



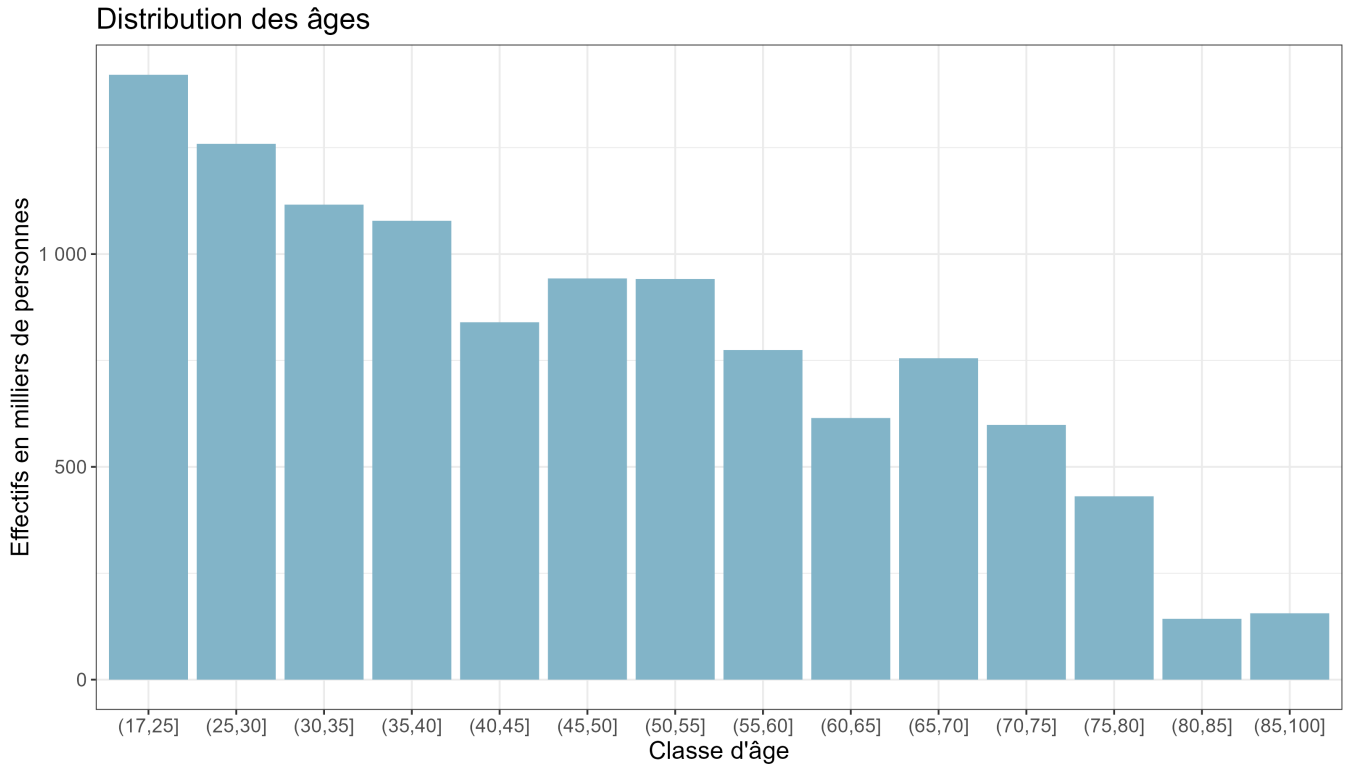
Les histogrammes

Construire une pyramide des âges avec des barres d'une largeur d'une année suppose par exemple une mesure précise de l'âge en années et un échantillon suffisamment grand pour éviter les creux (cf cours suivant sur l'échantillonnage).



La discrétisation

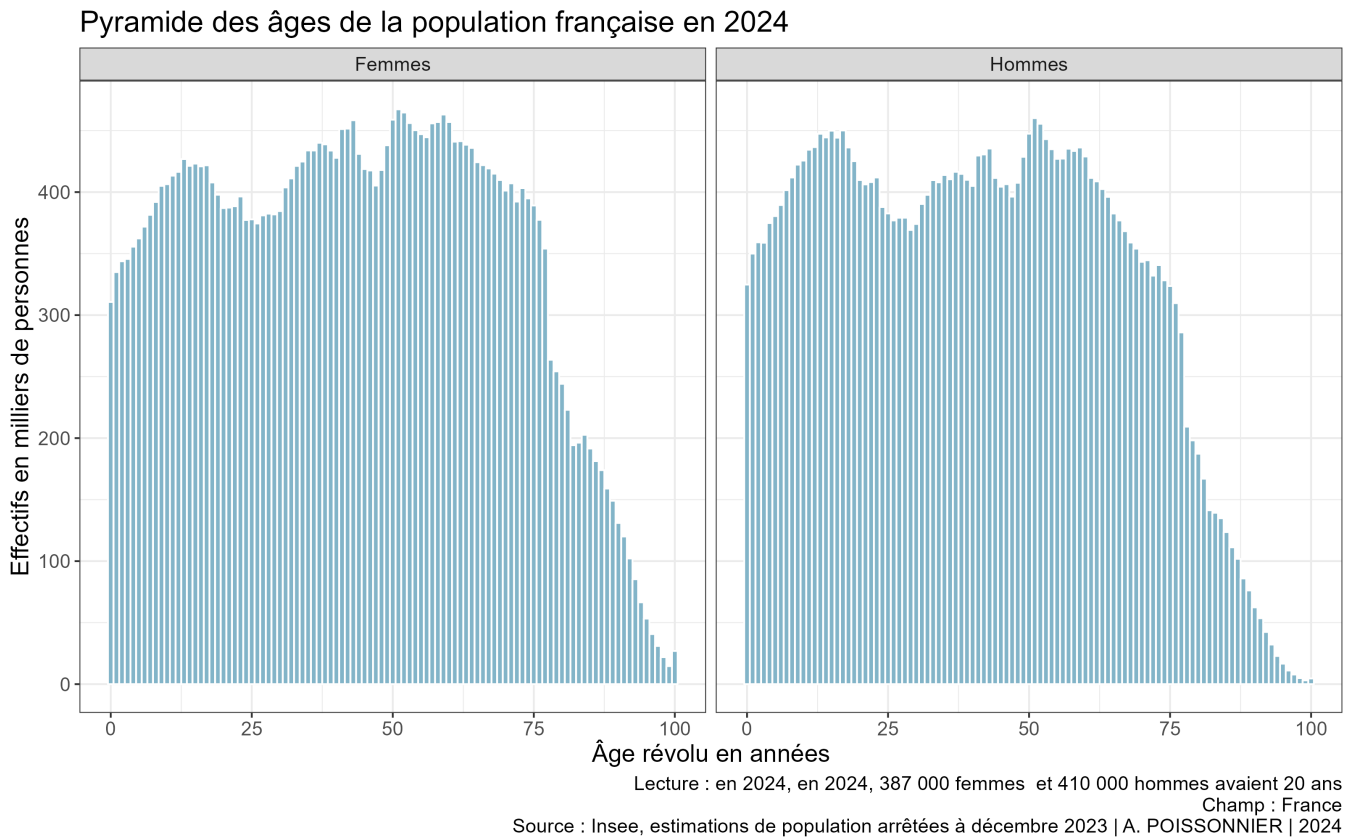
La *discrétisation* consiste à remplacer une mesure continue par une mesure discrète d'une dimension numérique. Les classes d'âges peuvent être construites par nécessité (manque de précision) ou par simplicité (logique d'équivalence de la catégorisation, où les différences d'unité ne sont pas significatives pour l'âge).



Lecture : en 2003, 1,4 millions de personnes étaient âgées de 18 à 25 ans
Champ : Individus de 18 ans et plus habitant en France métropolitaine
Source : Histoire de Vie 2003 (N = 2000) | A. POISSONNIER | 2024

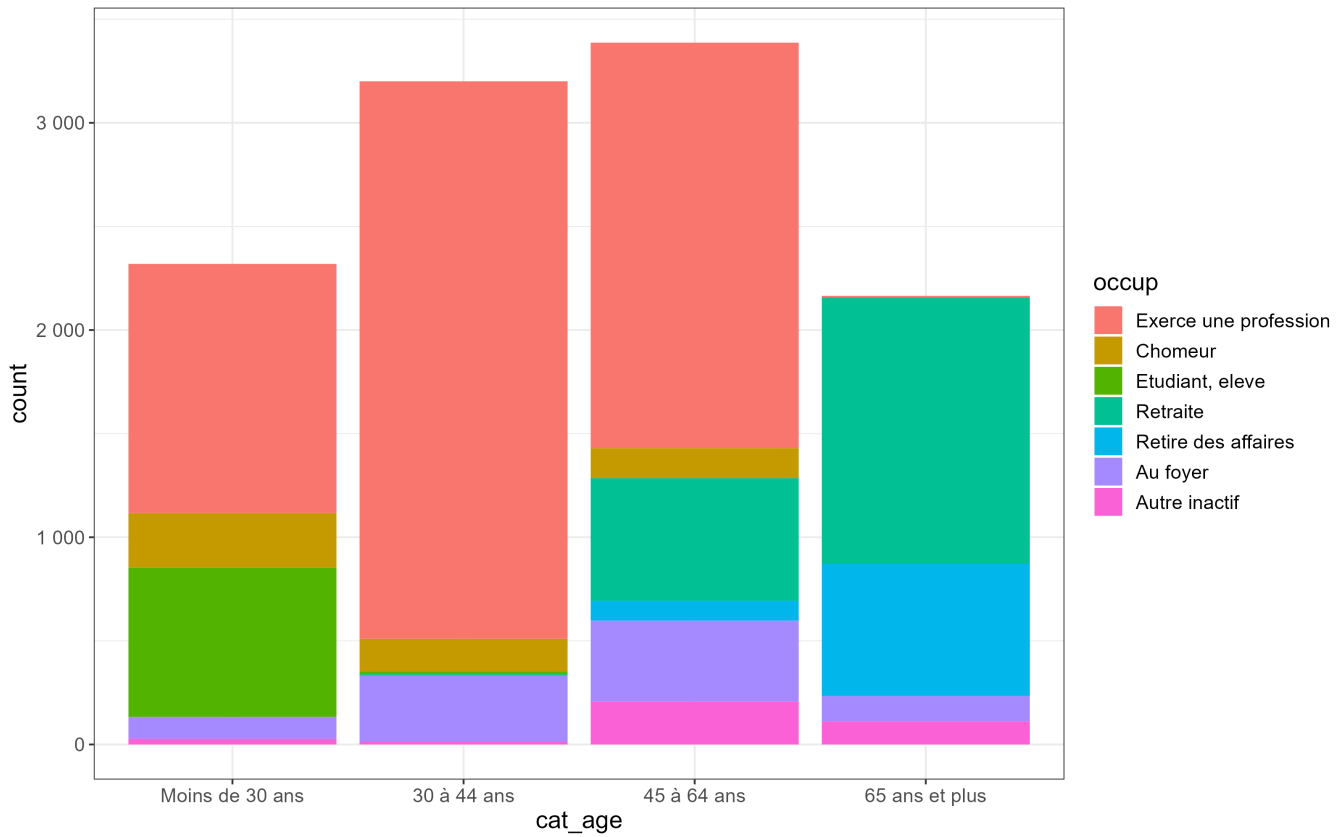
La comparaison côte à côte

Un même graphique peut être répliquer dans des sous-groupes distincts pour permettre la comparaison entre distributions mises côte à côte, la lisibilité n'étant pas garantie.



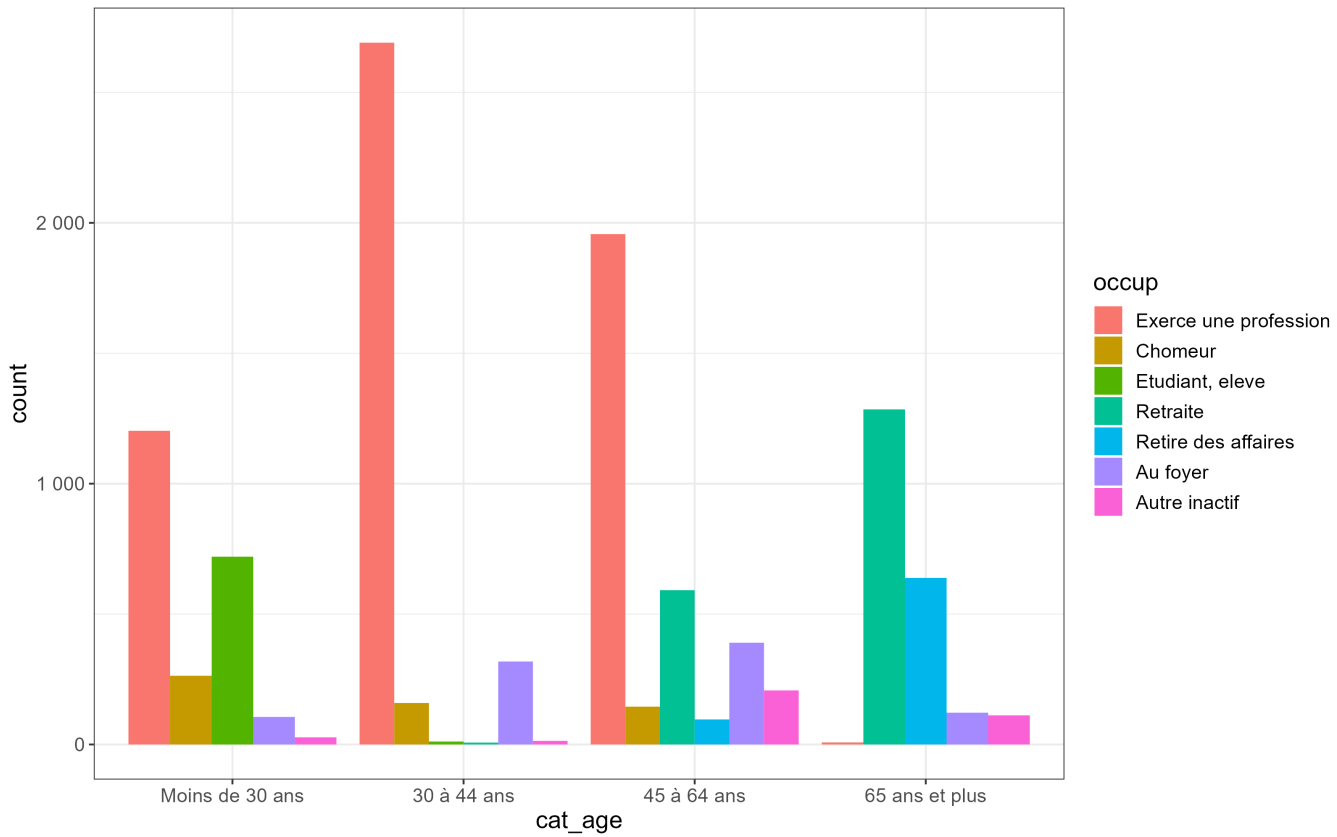
Croiser deux variables catégorielles

Un diagramme en barres peut permettre le croisement de deux variables catégorielles et la comparaison des distributions, par exemple en coloriant les parts respectives des modalités d'une variable au sein de sous-groupes définis par une deuxième variable.



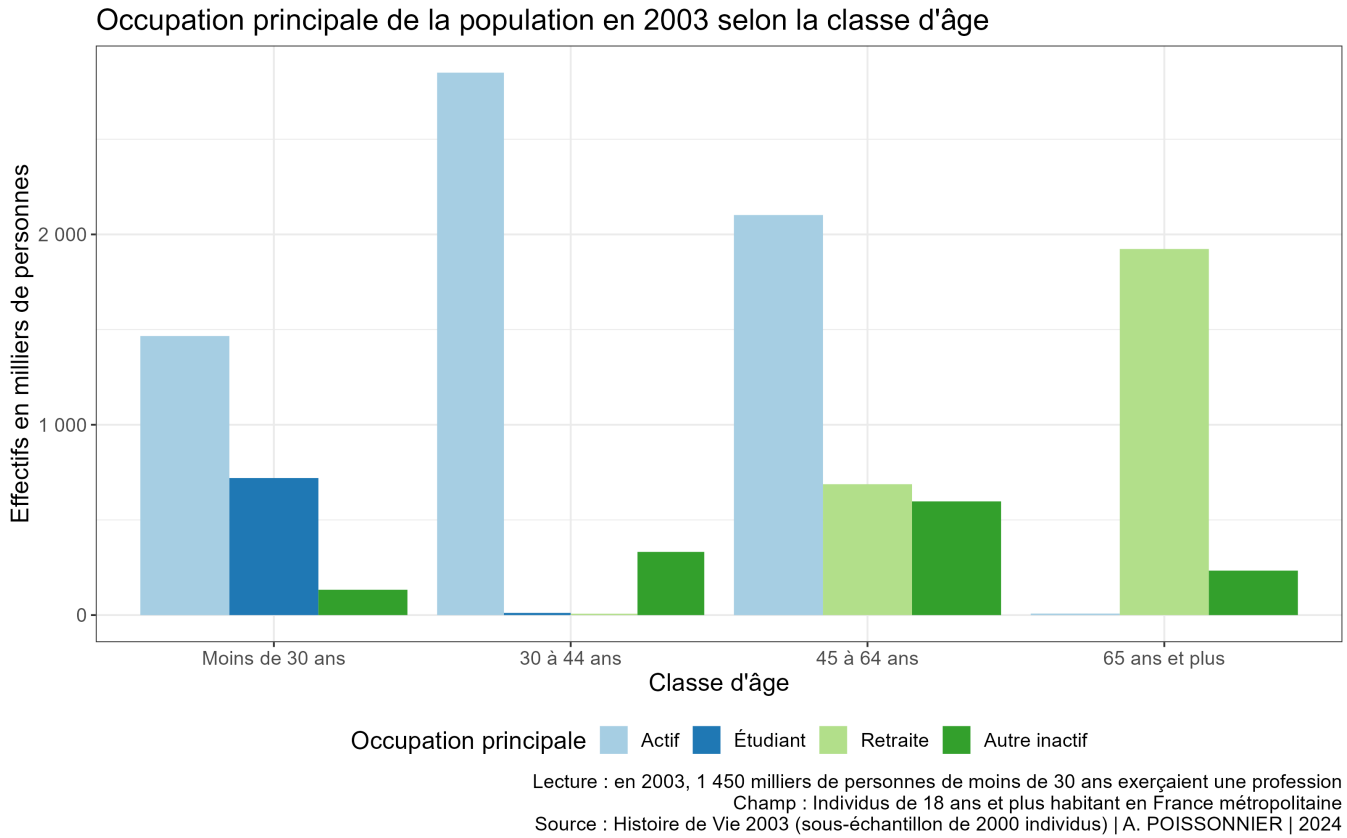
Croiser deux variables catégorielles

Un *diagramme en barres groupées* améliore la lisibilité d'une telle représentation et limite les erreurs (hauteur > longueur).
Comment peut-on rendre ce graphique plus synthétique ?



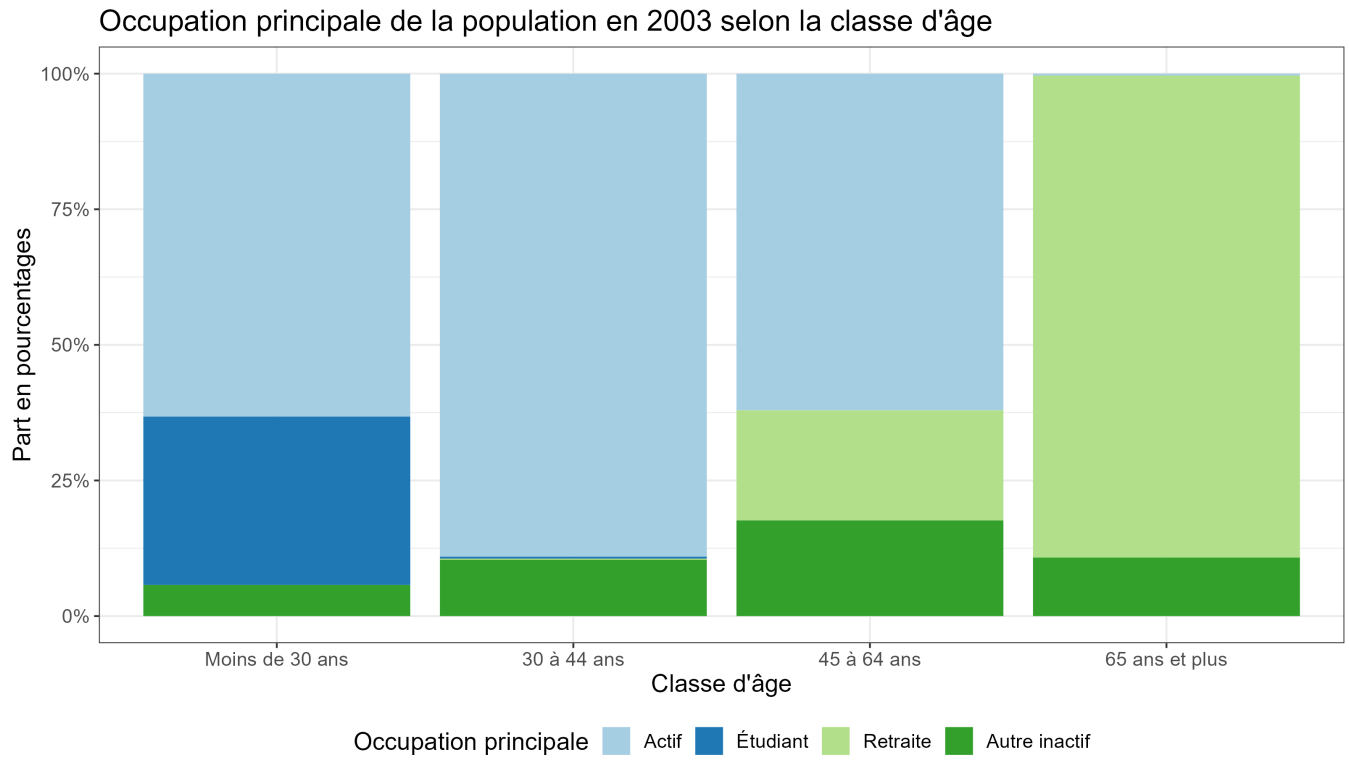
Croiser deux variables catégorielles

Un trop grand niveau de détail peut rendre la représentation trop complexe, et donc peu lisible. Le *recodage* consiste à renouveler l'opération de catégorisation pour simplifier l'information représentée.



Croiser deux variables catégorielles

Les parts plutôt que les effectifs peuvent aussi être représentées (ici, sans les mettre côte à côte pour se limiter à ce qui vous sera accessible en termes de maîtrise du code R).



Le tableau croisé associé

Le diagramme présenté est la traduction du tableau croisé de la classe d'âge avec l'occupation principale. Il doit contenir une ligne qui rappelle les parts dans l'ensemble de la population pour interpréter ces chiffres en termes de sur et sous-représentation.

Occupation principale de la population en 2003 selon la classe d'âge

	Actif	Étudiant	Retraite	Autre inactif
Moins de 30 ans	63%	31%	0%	6%
30 à 44 ans	89%	0%	0%	10%
45 à 64 ans	62%	0%	20%	18%
65 ans et plus	0%	0%	89%	11%
Ensemble	58%	7%	24%	11%

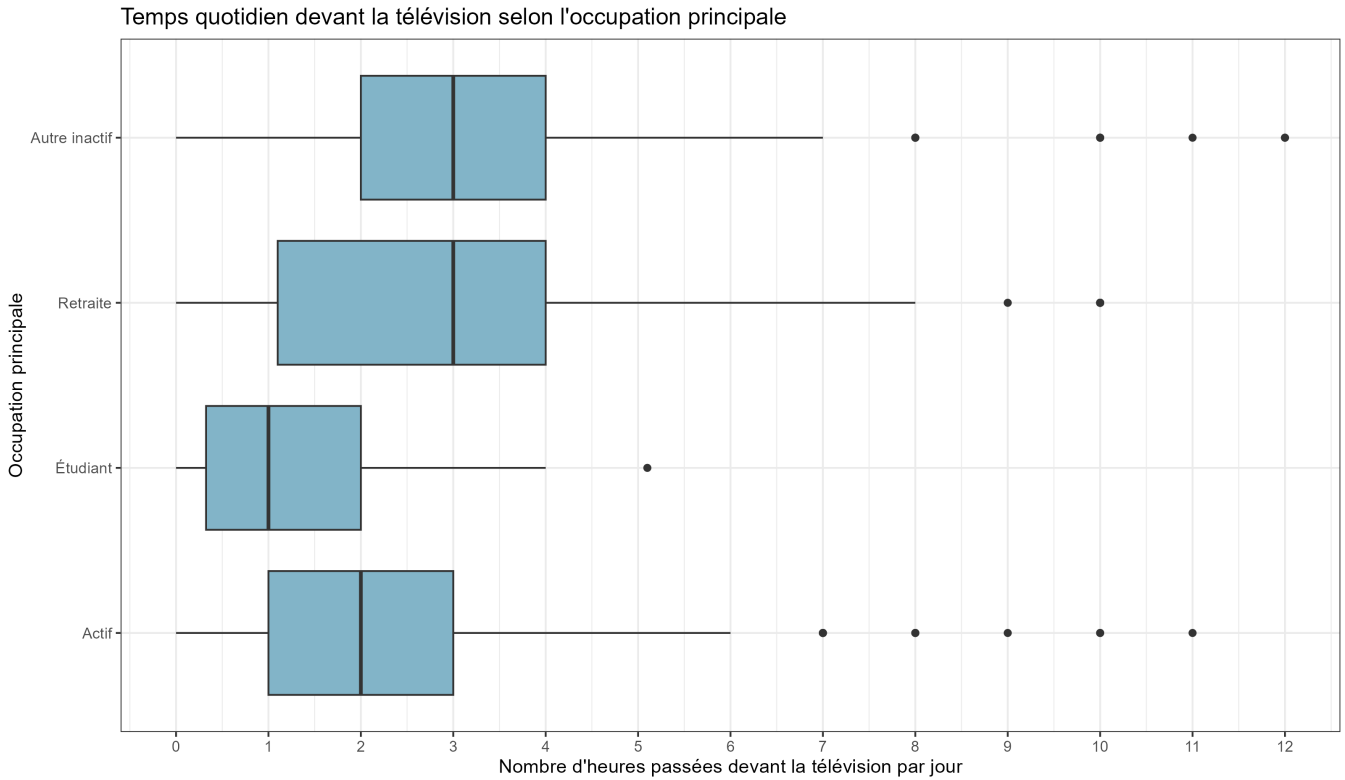
Lecture : en 2003, 63% des personnes de moins de 30 ans exerçaient une profession contre 58% dans l'ensemble de la population

Champ : Individus de 18 ans et plus habitant en France métropolitaine

Source : Histoire de Vie 2003 (sous-échantillon de 2000 individus) | A. POISSONNIER | 2024

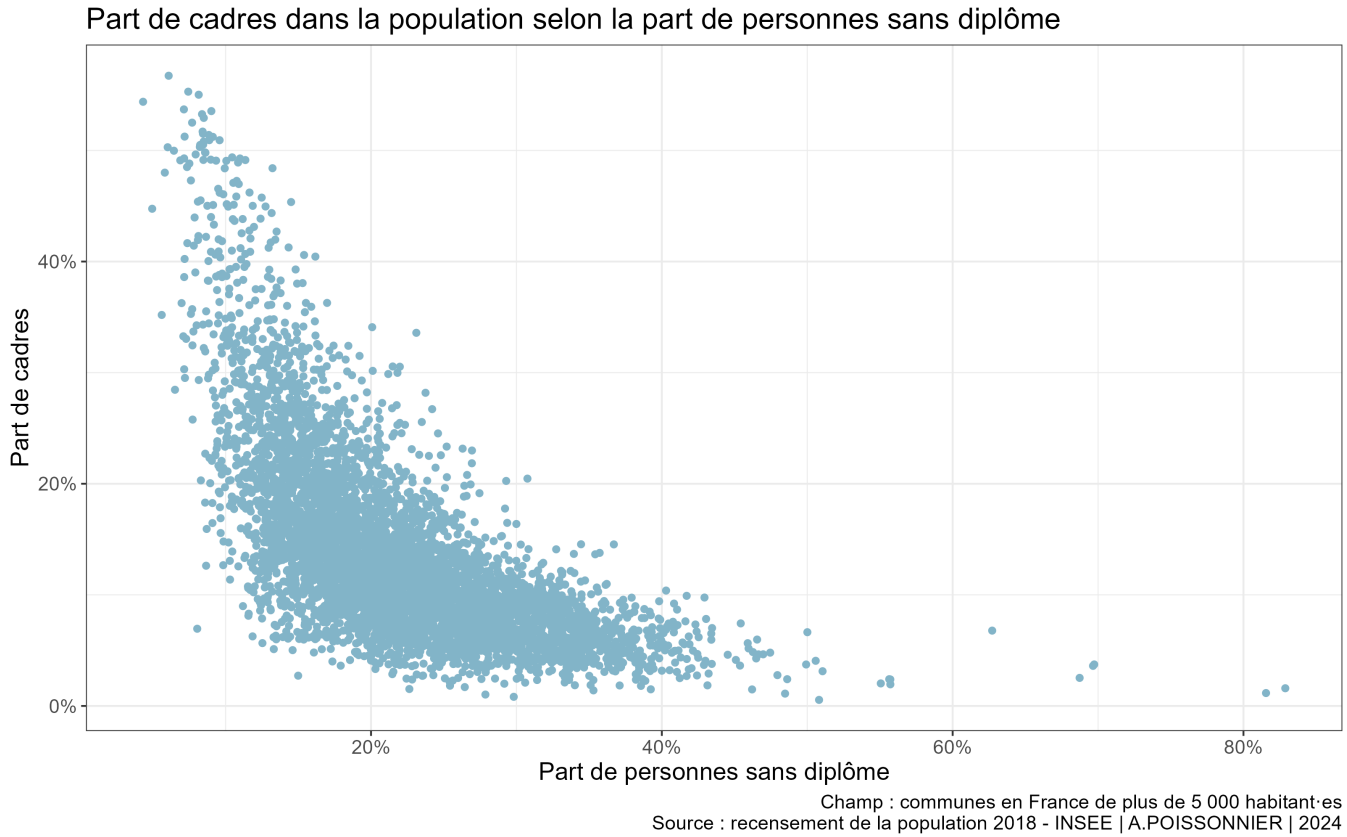
Croiser une variable catégorielle et une variable quantitative

Les *boîtes à moustaches* permettent de comparer la distribution d'une variable quantitative selon des sous-groupes définis par une variable catégorielle.



Croiser deux variables numériques

Les *nuages de points* permettent de croiser deux variables numériques, et donc de représenter la façon dont elles sont corrélées l'une à l'autre.



Suivre l'évolution d'un indicateur

Les *graphiques en courbes* permettent de suivre l'évolution d'un indicateur dans le temps. Chaque point correspond à une mesure et les segments qui les relient ne correspondent pas à des données réelles mais à une représentation simplifiée (linéaire) de l'évolution de l'indicateur entre deux mesures.