Rappels de Statistique

FX Jollois

TC - 2ème année - 2021/2022

Qu'est-ce que la statistique ?

- Ensemble de méthodes permettant de décrire et d'analyser des observations (communément appelées **données** de nos jours)
- Utilisé maintenant dans tous les secteurs d'activités
  - Economie et finance : marketing, sondages...
  - Industrie : fiabilité, contrôle qualité...
  - Santé : recherche médicale, gestion des hôpitaux...
  - Environnement : prévisions climatiques et météorologiques, pollution...
  - Web : réseaux, publicité...
  - ...
- Essor important avec le développement des outils informatiques et du web

Définitions de base

- **Population** : ensemble d'entités (personnes, objets...) étudiés
- **Individu** (ou *unité statistique*) : entité étudié
- **Variable** : caractéristique étudié sur chaque individu
- **Observation** : mesure
- **Série statistique** : série d'observations recueillies sur les individus
- **Tableau de données** : stockage de la série statistique
  - Individus croisant des variables
  - Chaque ligne représente un individu
  - Chaque colonne représente une variable (ou attribut)
  - C'est ce qu'on fait classiquement dans un tableur de type Excel

Données *Diamants*

- Exemple utilisé dans ce document
- ~54000 diamants (10 premières lignes ici)

carat	cut	color	clarity	depth	table	price	x	y	z
1.50	Premium	D	VS1	62.8	58	16793	7.29	7.26	4.57
1.30	Premium	I	IF	61.8	58	7760	7.00	6.95	4.31
1.66	Premium	D	SI1	62.0	59	14354	7.60	7.55	4.70
0.52	Ideal	G	VVS2	61.7	56	1911	5.14	5.16	3.18
0.25	Good	D	VVS2	64.6	56	462	3.97	3.99	2.57
0.31	Premium	D	SI1	61.7	59	732	4.35	4.31	2.67
0.32	Premium	E	VS2	60.5	58	702	4.44	4.48	2.70
0.31	Ideal	G	VVS1	60.6	57	816	4.38	4.40	2.66
1.01	Good	E	SI1	64.3	59	4106	6.28	6.31	4.05
0.70	Good	E	VS2	64.2	60	2394	5.58	5.64	3.60

Recensement vs Sondage

Recensement

Etude de tous les individus d'une population

- Recueil exhaustif de toutes les informations sur tous les entités
- Difficile à mettre en œuvre la plupart du temps

Sondage

Etude d'une partie de la population pour extrapolation sur l'ensemble de la population

- Partie des individus étudiés = **échantillon**
- Représentativité de l'échantillon ?

Type de variables

Variable quantitative

- Caractéristiques numériques : opérations de type somme ayant un sens
- **Continue** :
  - Mesurable
  - Ex : taille, poids, durée...
- **Discrète** :
  - Dénombrable ou mesurable en espace fini
  - Ex : âge, quantité en stock...

Type de variables

Variable qualitative

- Caractéristiques non numériques : opérations de type somme n'ayant pas de sens
  - Valeurs possibles : **Modalités** (ou catégories)
- **Nominale** :
  - Modalités n'ayant pas de lien entre elles
  - Ex : couleur des yeux, sexe...
  - Cas particulier : *binaire*
- **Ordinale** :
  - Modalités devant être triées dans un ordre spécifique
  - Ex : mois, sentiment...

Transformation de variable

Quantitative en qualitative

- Courant de transformer une variable **quantitative** en variable **qualitative ordinale**
  - Ex : Catégorie d'âge, Nombre d'enfants du foyer, ...
- Différents problèmes se posent
  - Combien de modalités (*intervalles* ici) ?
  - Taille identique des intervalles ou variable (*amplitude*) ?
  - Seuils des intervalles ?

Transformation de variable

Standardisation ou normalisation d'une variable quantitative

- Obligatoire pour l'utilisation de certaines méthodes statistiques
- 2 opérations sont réalisées :
  - Centrage : on retire la moyenne à chaque valeur
  - Réduction : on divise par la variance

$$x_{norm} = \frac{x - \bar{x}}{\sigma^2}$$

Premier problème : décrire les données

On parle de **Statistique descriptive** ou **exploratoire**

Objectifs

- Résumer l'information contenue dans les données
- Faire ressortir des éléments intéressants
- Poser des hypothèses sur des phénomènes potentiellement existant dans les données

Outils

- Description numérique (moyenne, occurrences, corrélation...)
- Description graphique (histogramme, diagramme en barres, nuage de points...)

Variable quantitative

- Moyenne $\bar{x}$

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

- Variance (et écart-type $\sigma(x)$)

$$\sigma^2(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Variable quantitative

- Médiane $med(x)$: valeur permettant de séparer les observations ordonnées prises par x en 2 groupes de même taille

$$med(x) = m | p(x \leq m) = .5$$

  - si n est impair : $med(x) = x_{(n+1)/2}$
  - si n est pair : $med(x) = \frac{x_{n/2} + x_{n/2+1}}{2}$
- Quantile $q_p(x)$: valeur pour laquelle une proportion p d'observations sont inférieures

$$q_p(x) = q | p(x \leq q) = p$$

  - Quartiles $Q1$ et $Q3$: respectivement 25% et 75% (utilisés dans les boîtes à moustaches)

Variable quantitative

Exemple : prix des diamants

Représentation numérique

Statistique	Valeur
Moyenne	3932.80
Ecart-Type	3989.44
Variance	15915629.42
Médiane	2401.00
Minimum	326.00
Maximum	18823.00

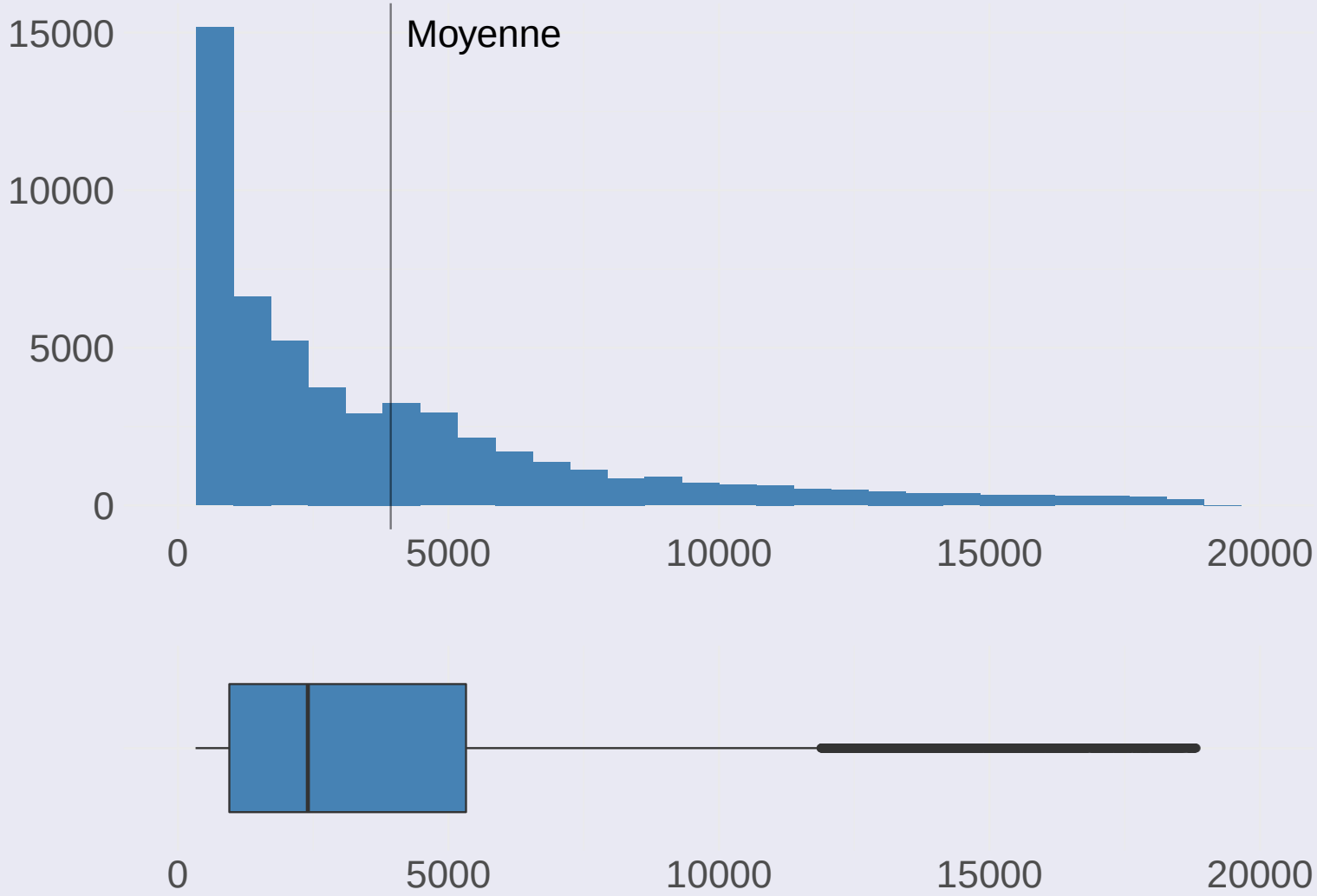
A regarder aussi :

- Divergence moyenne et médiane
  - Valeurs extrêmes présentes
  - Déséquilibre de la répartition des valeurs
- Présence de valeurs aberrantes
  - On parle d'**outliers**

Variable quantitative

Représentation graphique

Histogramme et boîte à moustaches



Variable qualitative

Nominale

- Modalités de la variable x : m_j (avec $j = 1, \ldots, p$)
- Effectif (ou occurrences) d'une modalité n_j : nombre d'individus ayant la modalité m_j
- Fréquence d'une modalité f_j

$$f_j = \frac{n_j}{n}$$

Ordinale

- Effectif cumulé n_j^{cum} : nombre d'individus ayant une modalité entre n_1 et n_j
- Fréquence cumulée

$$n_j^{cum} = \sum_{k=1}^{j} n_k \text{ and } f_j^{cum} = \sum_{k=1}^{j} f_k$$

Variable qualitative

Exemple : Qualité de découpe (*ordinale* de plus)

Représentation numérique

Modalités	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
Fair	1610	1610	0.03	0.03
Good	4906	6516	0.09	0.12
Very Good	12082	18598	0.22	0.34
Premium	13791	32389	0.26	0.60
Ideal	21551	53940	0.40	1.00

A regarder aussi :

- Différence entre les proportions
- Si modalités peu fréquentes, regroupement de modalités à envisager
  - Attention au sens de ces regroupements

Variable qualitative

Représentation graphique

Diagramme en barres

Quantitative vs quantitative

- Covariance

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

  - Problème : non bornée et donc non exploitable
- Coefficient de corrélation linéaire (de *Pearson*)

$$\rho(x, y) = \frac{cov(x, y)}{\sigma^2(x)\sigma^2(y)}$$

  - Covariance des variables normalisées
  - Valeurs comprises entre -1 et 1
    - 0 : pas de lien linéaire (autre type de lien possible)
    - 1 : lien positif fort (si x augmente, y augmente)
    - -1 : lien négatif fort (si x augmente, y diminue)

Quantitative vs quantitative

Exemple : Prix et Carat

Représentation numérique

Statistique	Valeur
Covariance	1742.77
Corrélation	0.92

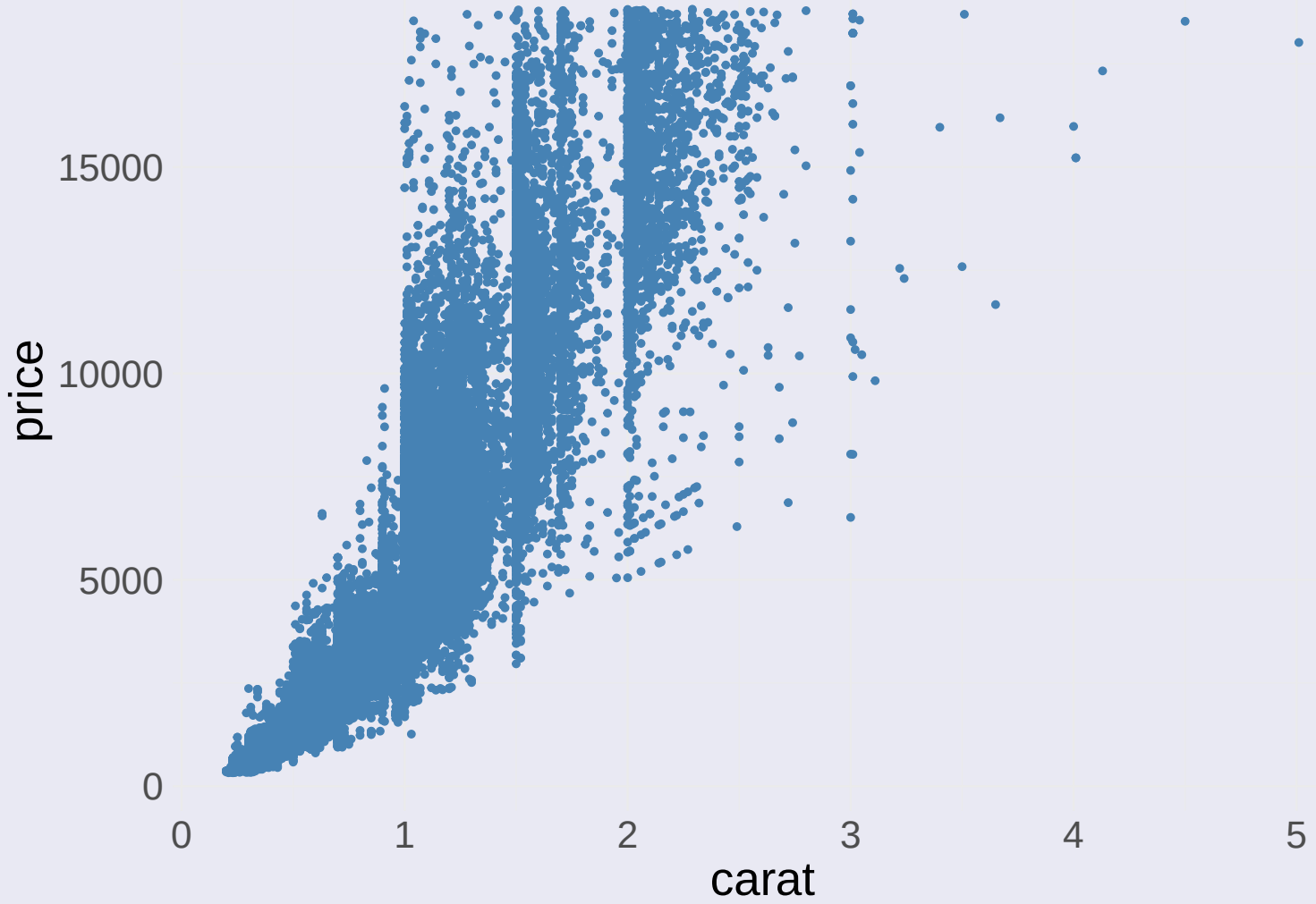
A regarder aussi :

- Présence d'**outliers** avec un comportement atypique

Quantitative vs quantitative

Représentation graphique

Nuage de points



Anscombe

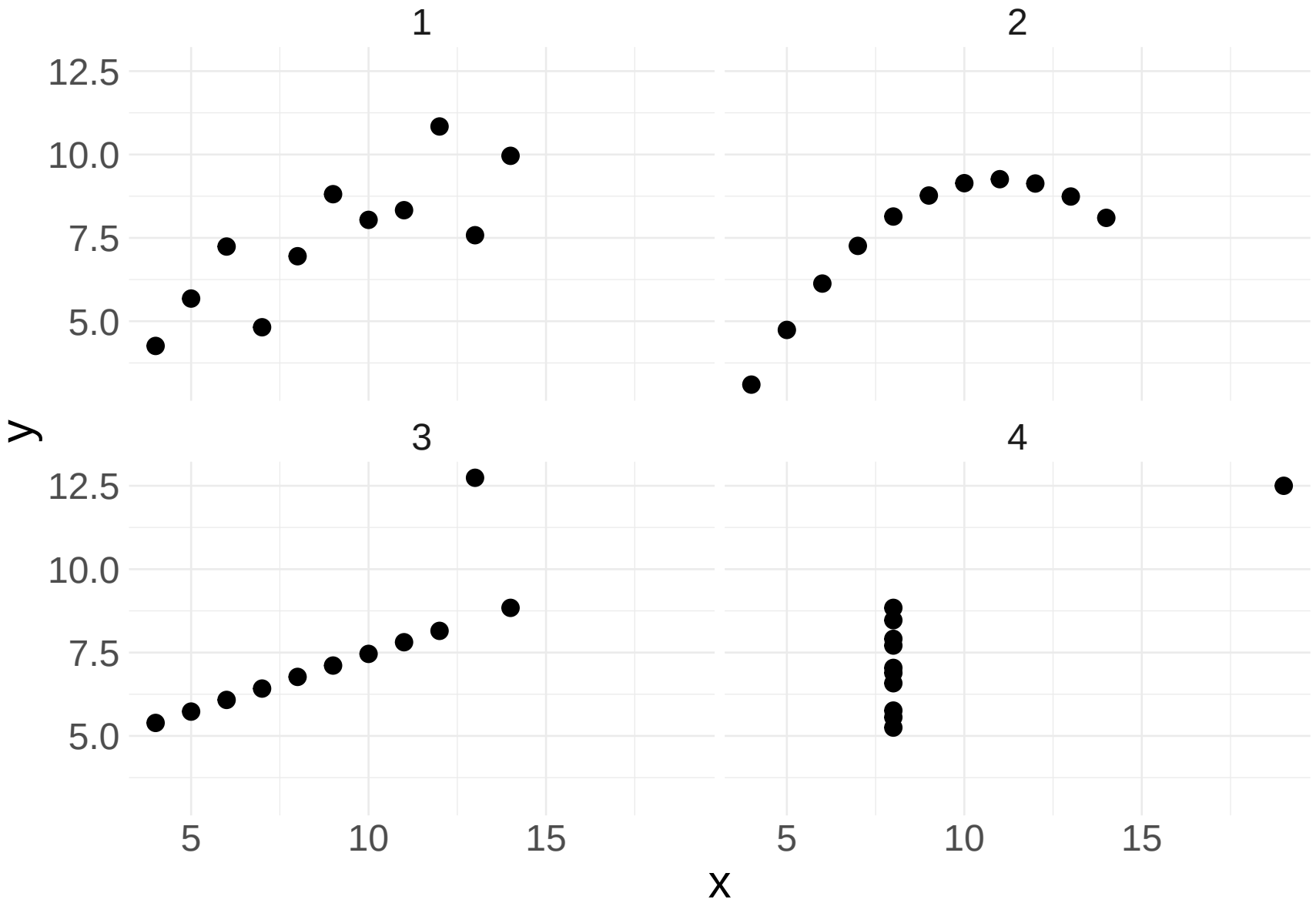
La visualisation est aussi importante (voire plus) que la représentation numérique !

Entre ces quatres séries :

- même moyenne et même variance pour x et y
- même coefficient de corrélation entre les deux

	1	2	3	4
Moyenne(x)	9.00	9.00	9.00	9.00
Moyenne(y)	7.50	7.50	7.50	7.50
Ecart-type(x)	3.32	3.32	3.32	3.32
Ecart-type(y)	2.03	2.03	2.03	2.03
Covariance	5.50	5.50	5.50	5.50
Corrélation	0.82	0.82	0.82	0.82

Anscombe



Qualitative vs qualitative

- Table de contingence
  - Croisement des 2 ensembles de modalités, avec le nombre d'individus ayant chaque couple de modalités
  - n_{ij} : Nombre d'observations ayant la modalité i pour x et j pour y
  - $n_{i.}$: Effectif marginal (nombre d'observations ayant la modalité i pour x)
  - $n_{.j}$: Effectif marginal (nombre d'observations ayant la modalité j pour y)

	1	...	j	...	l	total
1						
...						
i			n_ij			n_i.
...						
k						
total			n_.j			n_.. = n

Qualitative vs qualitative

- Profils lignes et colonnes
  - Distribution d'une variable conditionnellement aux modalités de l'autre

Qualitative vs qualitative

Exemple : Qualité et couleur

Représentation numérique

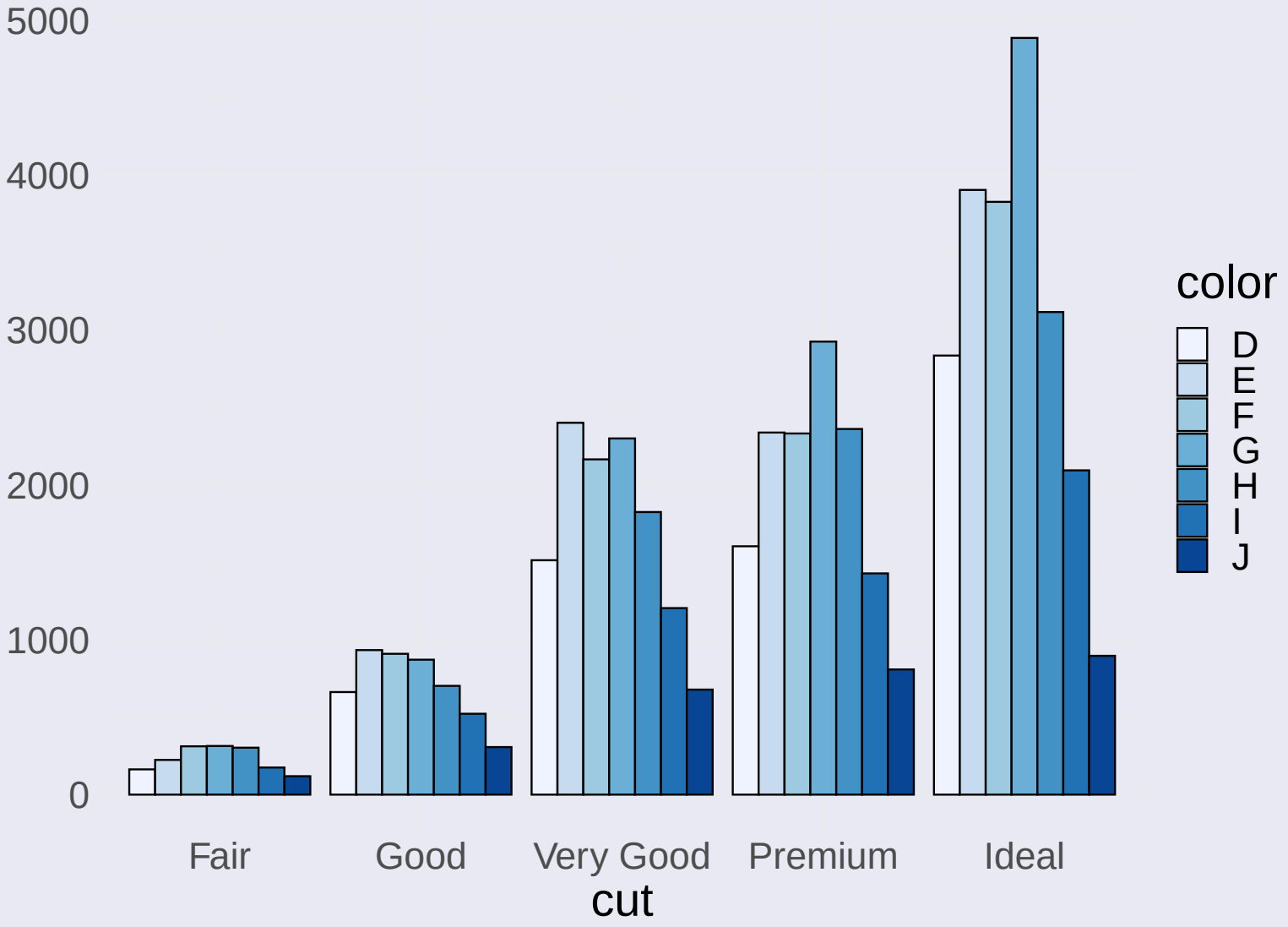
color	Fair	Good	Very Good	Premium	Ideal
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

A regarder aussi :

- Couple de modalités très peu pris
- Ici aussi, regroupement de modalités à envisager éventuellement

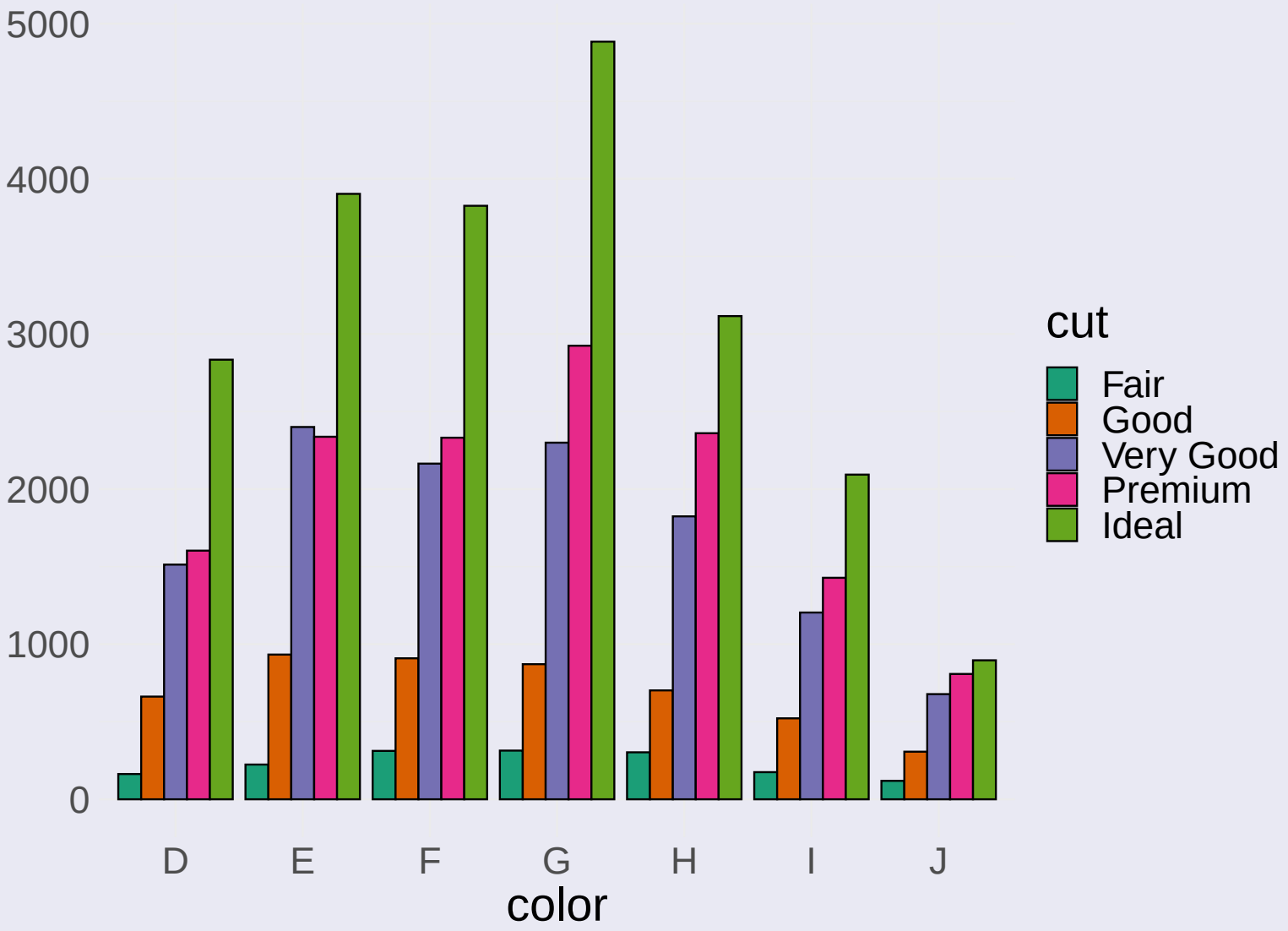
Qualitative vs qualitative

Représentation graphique



Qualitative vs qualitative

Représentation graphique



Qualitative vs qualitative

Exemple : Qualité et couleur

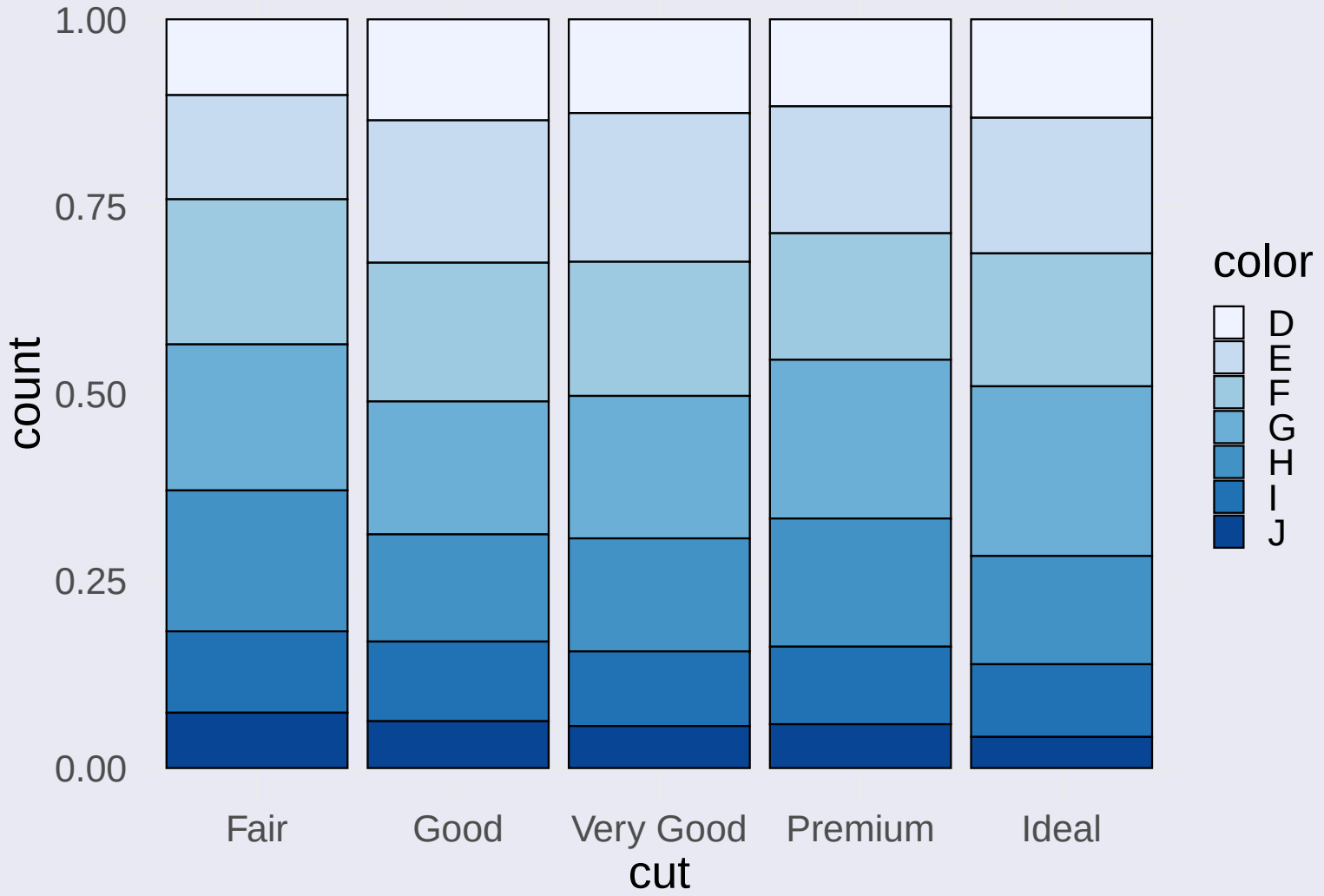
Représentation numérique

Profils colonnes ici (sommes en colonnes = 100%)

color	Fair	Good	Very Good	Premium	Ideal
D	0.10	0.13	0.13	0.12	0.13
E	0.14	0.19	0.20	0.17	0.18
F	0.19	0.19	0.18	0.17	0.18
G	0.20	0.18	0.19	0.21	0.23
H	0.19	0.14	0.15	0.17	0.14
I	0.11	0.11	0.10	0.10	0.10
J	0.07	0.06	0.06	0.06	0.04

Qualitative vs qualitative

Représentation graphique



Qualitative vs qualitative

Exemple : Qualité et couleur

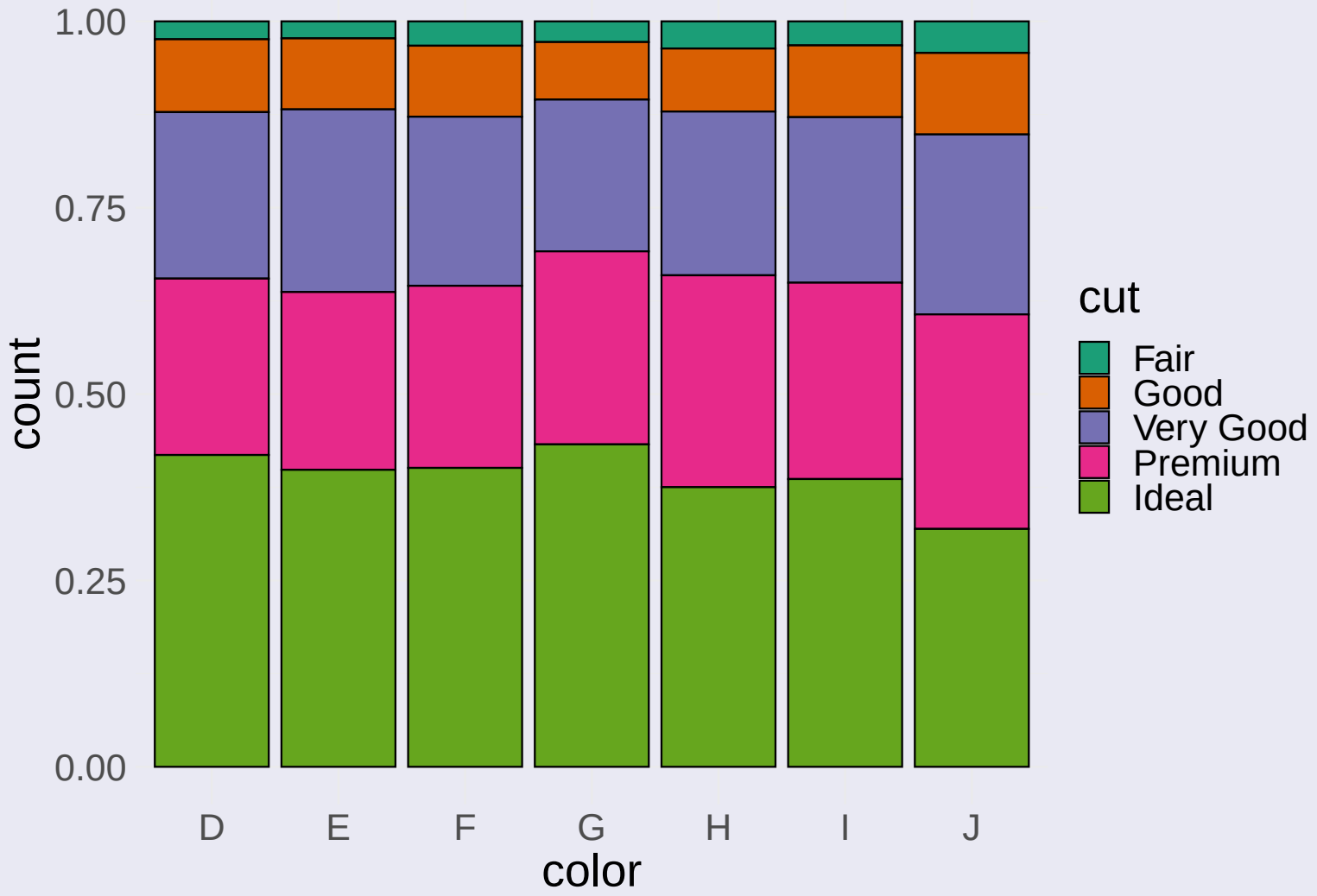
Représentation numérique

Profils lignes ici (sommes en lignes = 100%)

color	Fair	Good	Very Good	Premium	Ideal
D	0.02	0.10	0.22	0.24	0.42
E	0.02	0.10	0.24	0.24	0.40
F	0.03	0.10	0.23	0.24	0.40
G	0.03	0.08	0.20	0.26	0.43
H	0.04	0.08	0.22	0.28	0.38
I	0.03	0.10	0.22	0.26	0.39
J	0.04	0.11	0.24	0.29	0.32

Qualitative vs qualitative

Représentation graphique



Qualitative vs quantitative

- Soit Y la variable qualitative à m modalités, et X la variable quantitative
- Sous-populations déterminées par les modalités de Y
- Indicateurs calculés pour chaque modalité

$$\bar{x_j} = \frac{1}{n_j} \sum_{i|y_i=j} x_i$$

$$\sigma^2(x_j) = \frac{1}{n_j} \sum_{i|y_i=j} (x_i - \bar{x_j})^2$$

Qualitative vs quantitative

Exemple : Qualité et prix

Représentation numérique

cut	Moyenne	Ecart-type	Médiane
Fair	4358.76	3560.39	3282.0
Good	3928.86	3681.59	3050.5
Very Good	3981.76	3935.86	2648.0
Premium	4584.26	4349.20	3185.0
Ideal	3457.54	3808.40	1810.0

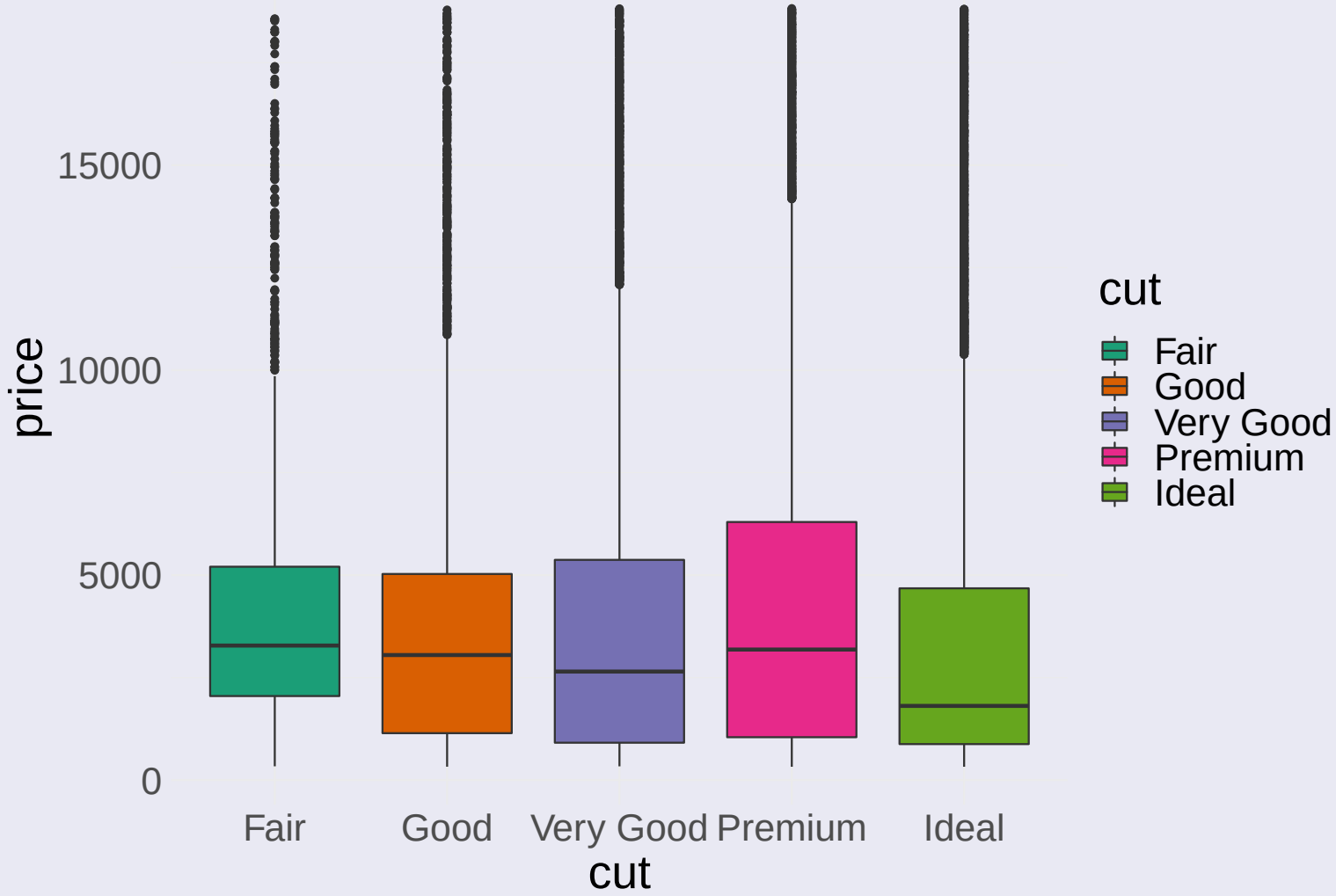
A regarder aussi :

- Outliers

Qualitative vs quantitative

Représentation graphique

Boîte à moustaches



Deuxième problème : Extrapoler à partir de données

On parle alors de **statistique inférentielle**

Cadre

- Données issues d'un échantillon d'une population
- Modèle probabiliste sur la population
- Méthodes d'échantillonnage pour choisir au mieux l'échantillon

Objectifs

- Etendre les conclusions faites sur l'échantillon à toute la population
- Valider (ou non) des hypothèses faites sur la population en analysant l'échantillon

Outils

- Estimation : approximer des paramètres de la population à partir de l'échantillon
- Test : valider les hypothèses
- Modélisation : rechercher des liens entre variables