Rappels de Statistique

FX Jollois

BUT TC - 2ème année

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

Données `tips`

Serveur notant des infos sur chaque table dont le pourboire

- Exemple utilisé dans ce document
- 10 premières lignes

total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.50	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4
25.29	4.71	Male	No	Sun	Dinner	4
8.77	2.00	Male	No	Sun	Dinner	2
26.88	3.12	Male	No	Sun	Dinner	4
15.04	1.96	Male	No	Sun	Dinner	2
14.78	3.23	Male	No	Sun	Dinner	2

Définitions de base

- **Population** : ensemble d'entités (personnes, objets, . . .) étudiées
- **Individu** (ou *unité statistique*) : entité étudiée
- **Variable** : caractéristique étudiée 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

Recensement vs Sondage

2 méthodes de recueil de données

Recensement

Etude de tous les individus d'une population

- Recueil exhaustif de toutes les informations sur toutes 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* : 2 valeurs possibles uniquement (Ex : oui/non, présence/absence...)

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)
  - Quantiles usuels : .01 (1%), .1 (10%), .9 (90%) et .99 (99%)

Variable quantitative

Exemple : montant payé par table

Représentation numérique

Statistique	Valeur
Moyenne	19.79
Ecart-Type	8.90
Variance	79.25
Médiane	17.80
Minimum	3.07
Maximum	50.81

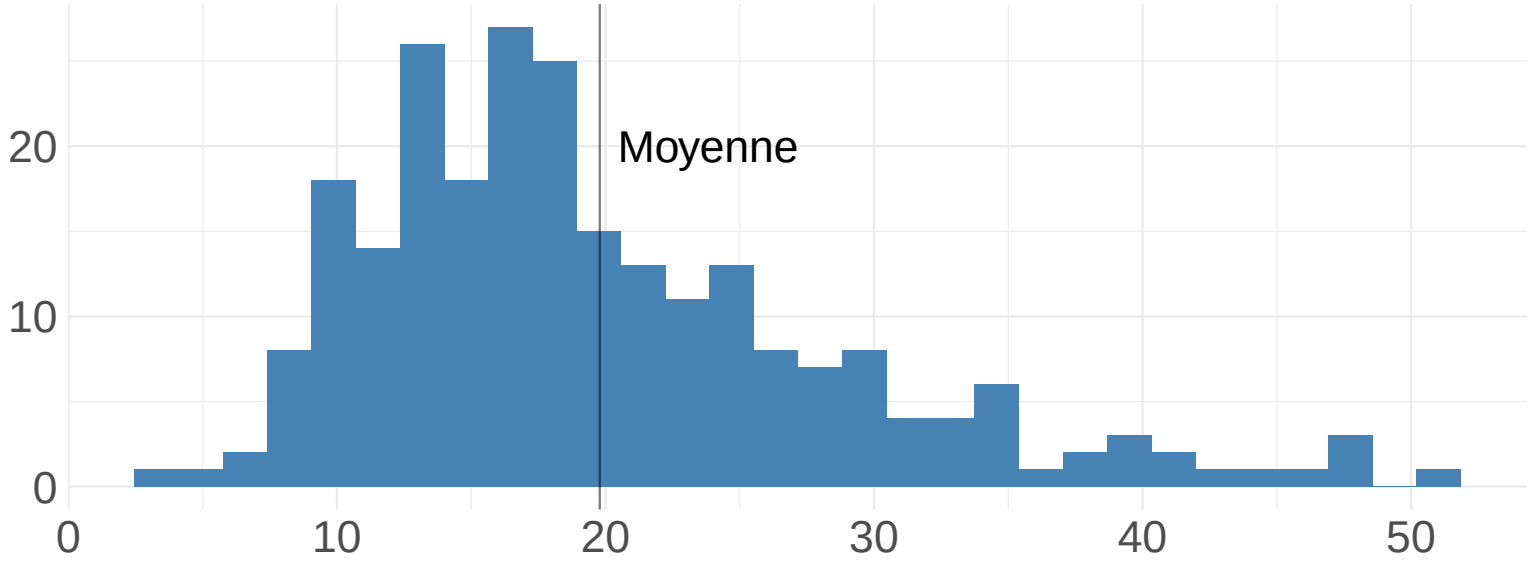
A regarder aussi :

- Si divergence moyenne et médiane, valeurs extrêmes présentes
  - Déséquilibre de la répartition des valeurs
- Présence de valeurs aberrantes (nommés **outliers**)

Variable quantitative

Représentation graphique

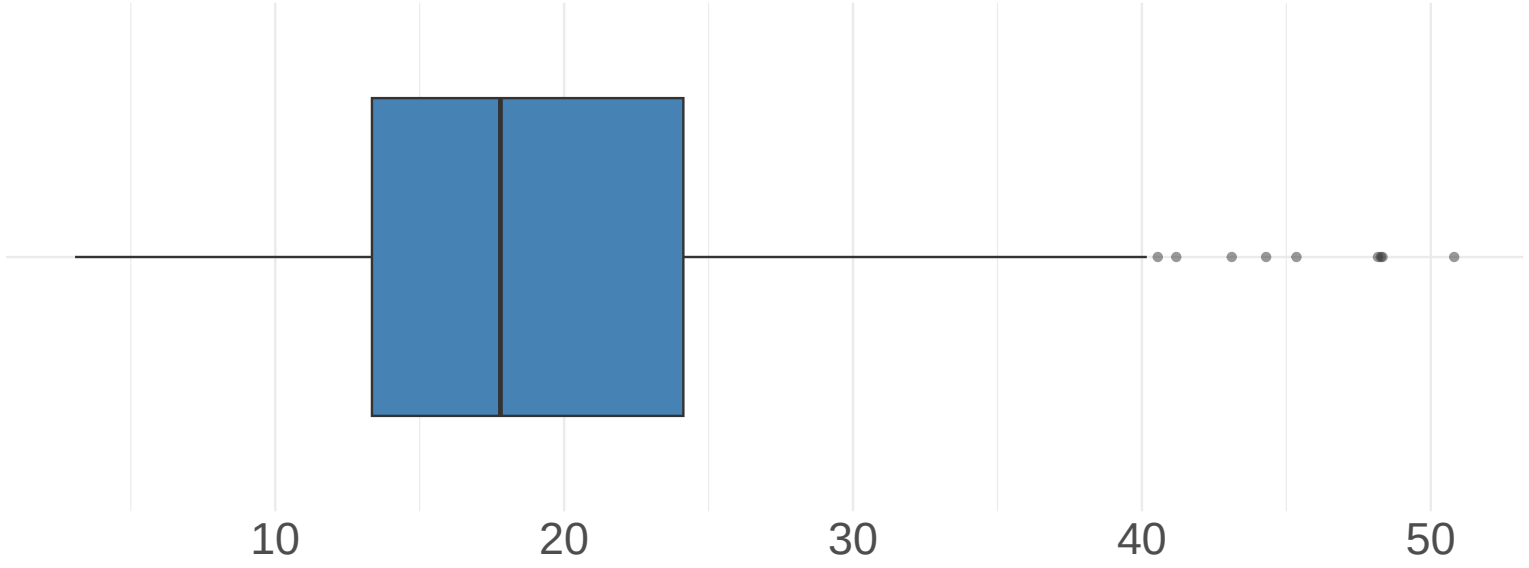
Histogramme



Variable quantitative

Représentation graphique

Boîte à moustaches



Variable qualitative

Nominale

- Modalités de la variable x : m_j (avec $j = 1, ..., p$)
- 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 : Jour de la semaine (*ordinale* de plus)

Représentation numérique

Modalités	Effectifs	Eff. cum.	Fréquences	Fréq. cum.
Fri	19	19	0.08	0.08
Sat	87	106	0.36	0.43
Sun	76	182	0.31	0.75
Thur	62	244	0.25	1.00

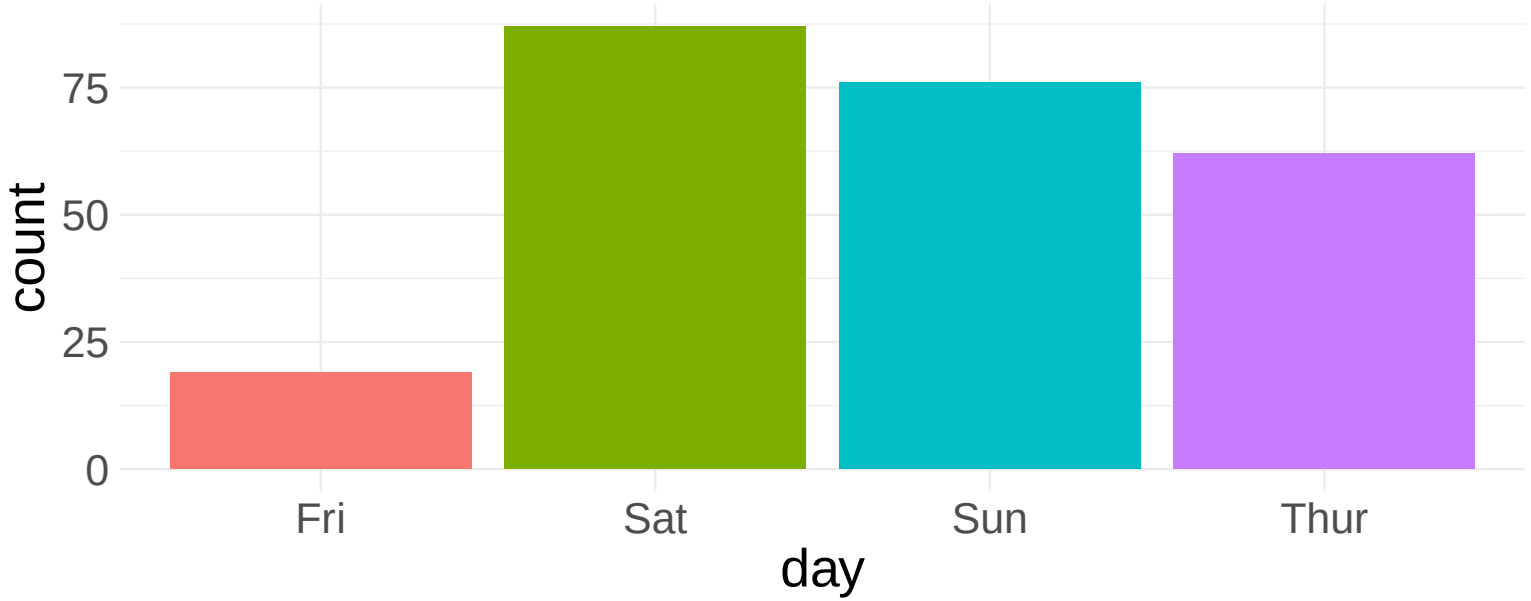
A regarder aussi :

- Différence entre les proportions
- Si modalités peu fréquentes, regroupement de modalités à envisager

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 : Montant de la table et Pourboire

Représentation numérique

Statistique	Valeur
Covariance	8.32
Corrélation	0.68

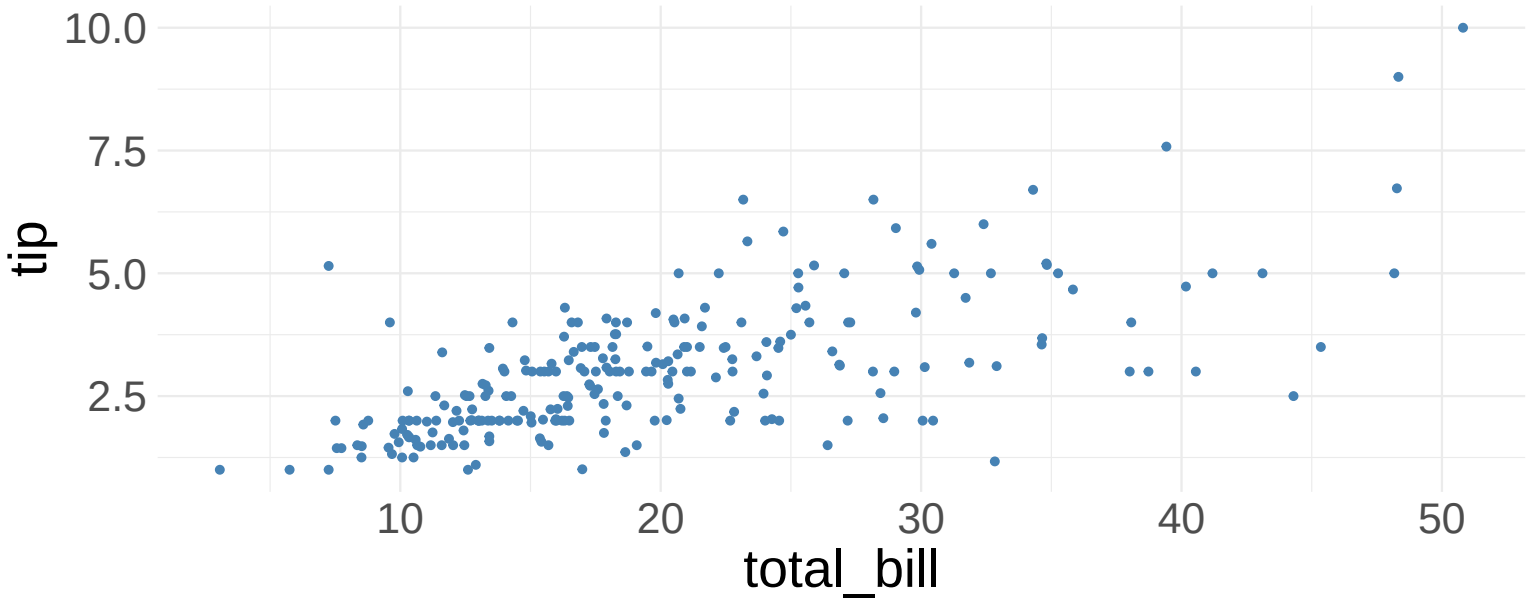
A regarder aussi :

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

Quantitative vs quantitative

Représentation graphique

Nuage de points



Anscombe

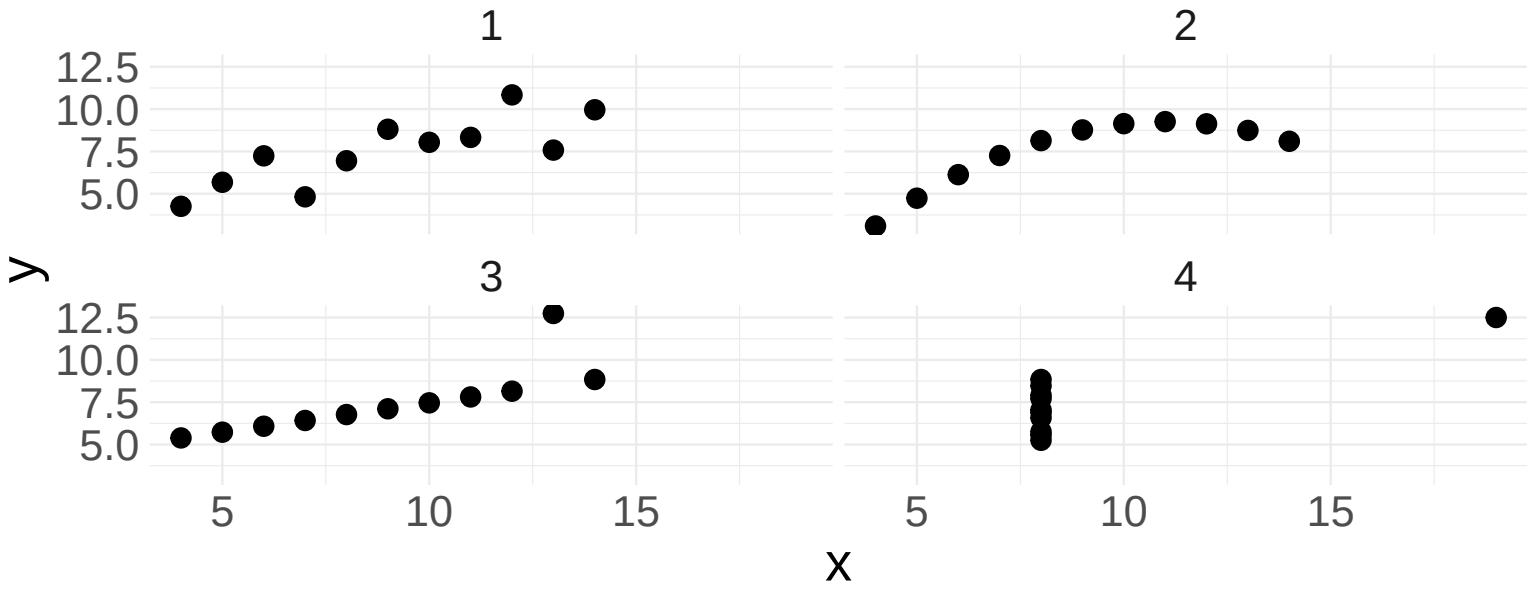
La visualisation est aussi importante (voire plus) que la représentation numérique !

Entre ces quatre 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
- Profil ligne
  - Pour une ligne i : $\frac{n_{ij}}{n_{i.}}$
  - Somme des valeurs en lignes $= 100\%$

-Profil colonne - Pour une colonne j : $\frac{n_{ij}}{n_{.j}}$ - Somme des valeurs en colonnes $= 100\%$

Qualitative vs qualitative

Exemple : Jour de la semaine et Présence de fumeur

Représentation numérique

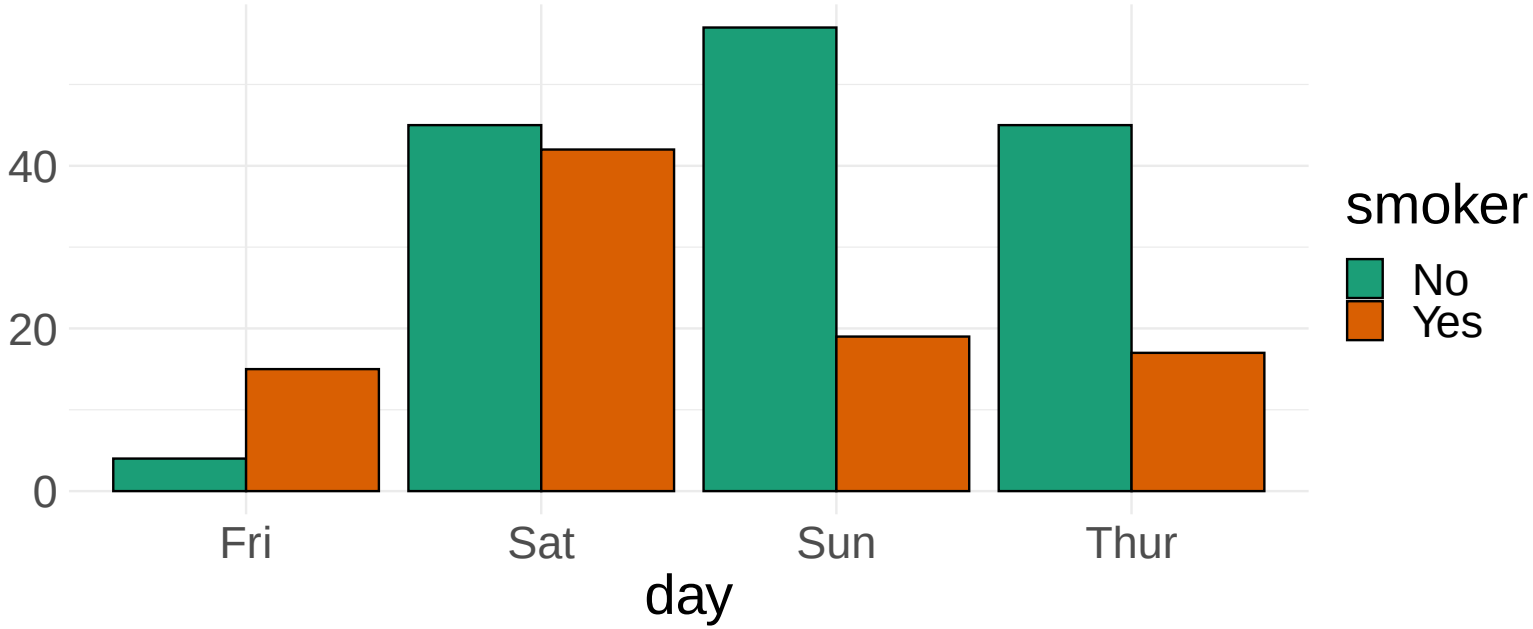
color	No	Yes
Fri	4	15
Sat	45	42
Sun	57	19
Thur	45	17

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

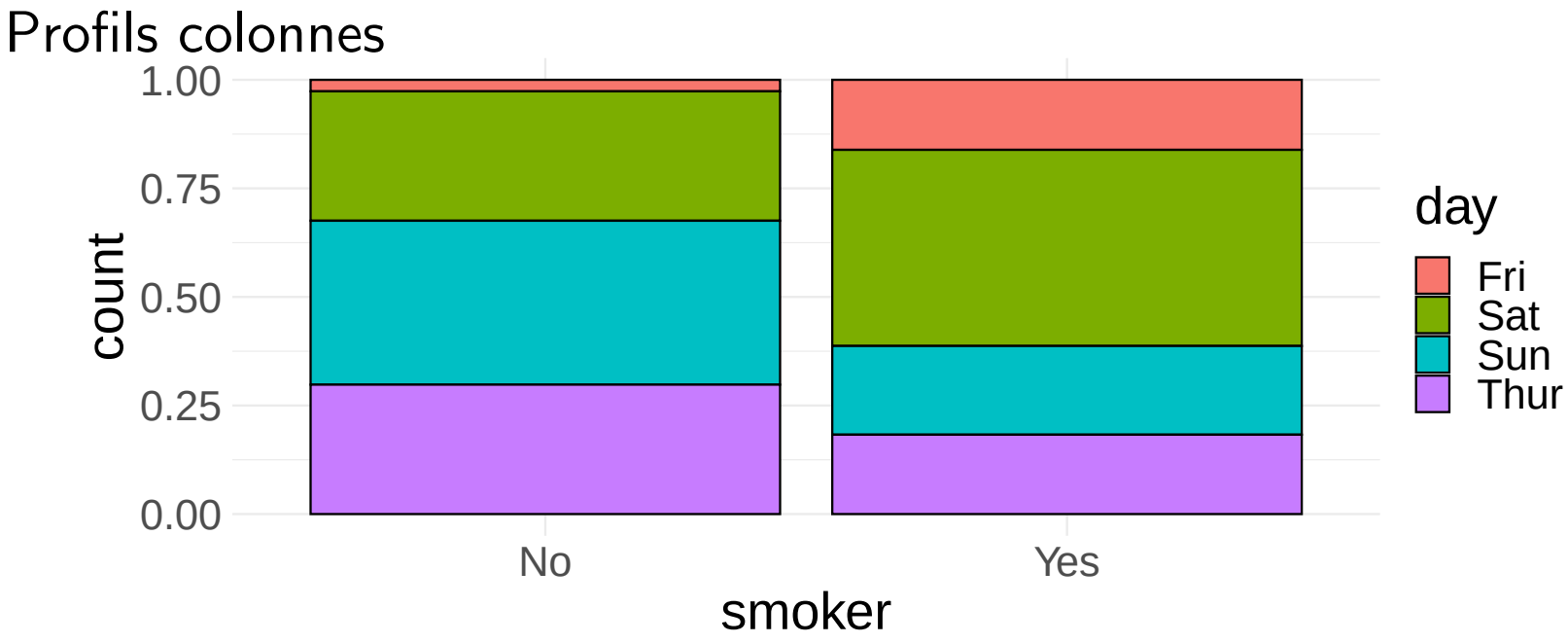
Représentation numérique

Profils colonnes ici (sommes en colonnes = 100%)

color	No	Yes
Fri	0.03	0.16
Sat	0.30	0.45
Sun	0.38	0.20
Thur	0.30	0.18

Qualitative vs qualitative

Représentation graphique



Qualitative vs qualitative

Représentation numérique

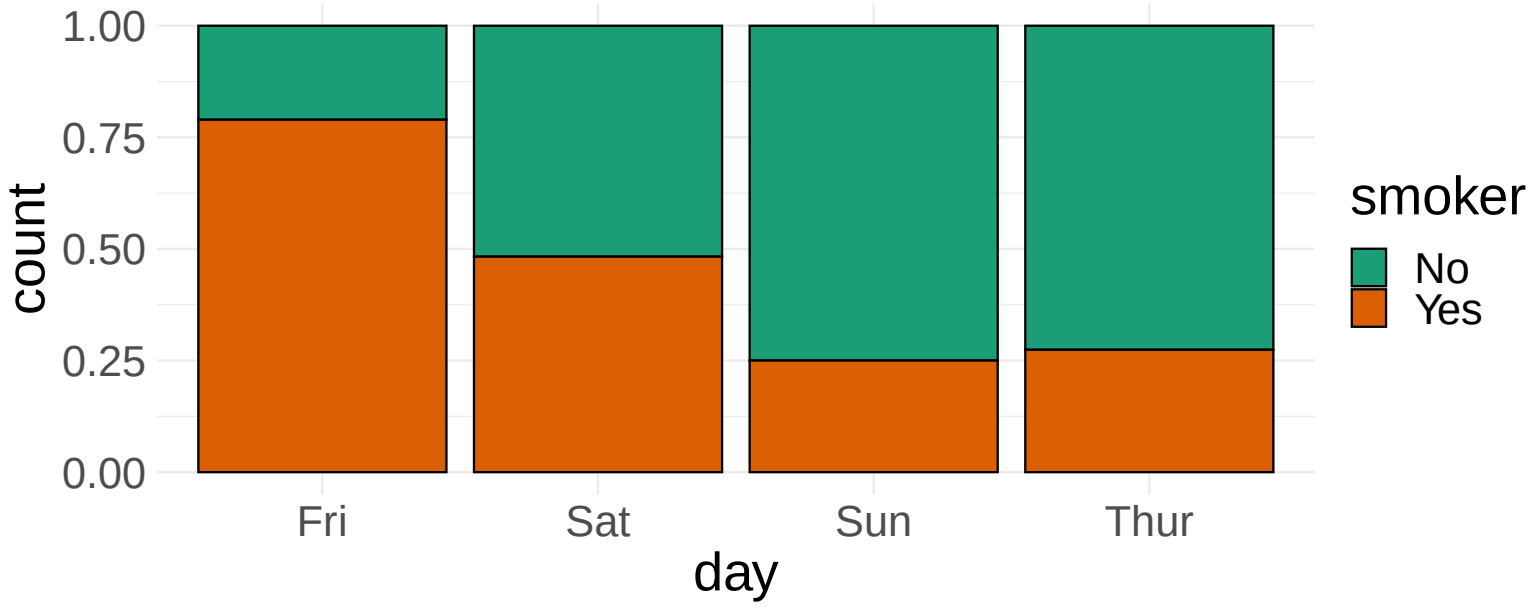
Profils lignes ici (sommes en lignes = 100%)

color	No	Yes
Fri	0.21	0.79
Sat	0.52	0.48
Sun	0.75	0.25
Thur	0.73	0.27

Qualitative vs qualitative

Représentation graphique

Profils lignes



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 k

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

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

Qualitative vs quantitative

Exemple : Montant payé et Jour de la semaine

Représentation numérique

day	Moyenne	Ecart-type	Médiane
Fri	17.15	8.30	15.38
Sat	20.44	9.48	18.24
Sun	21.41	8.83	19.63
Thur	17.68	7.89	16.20

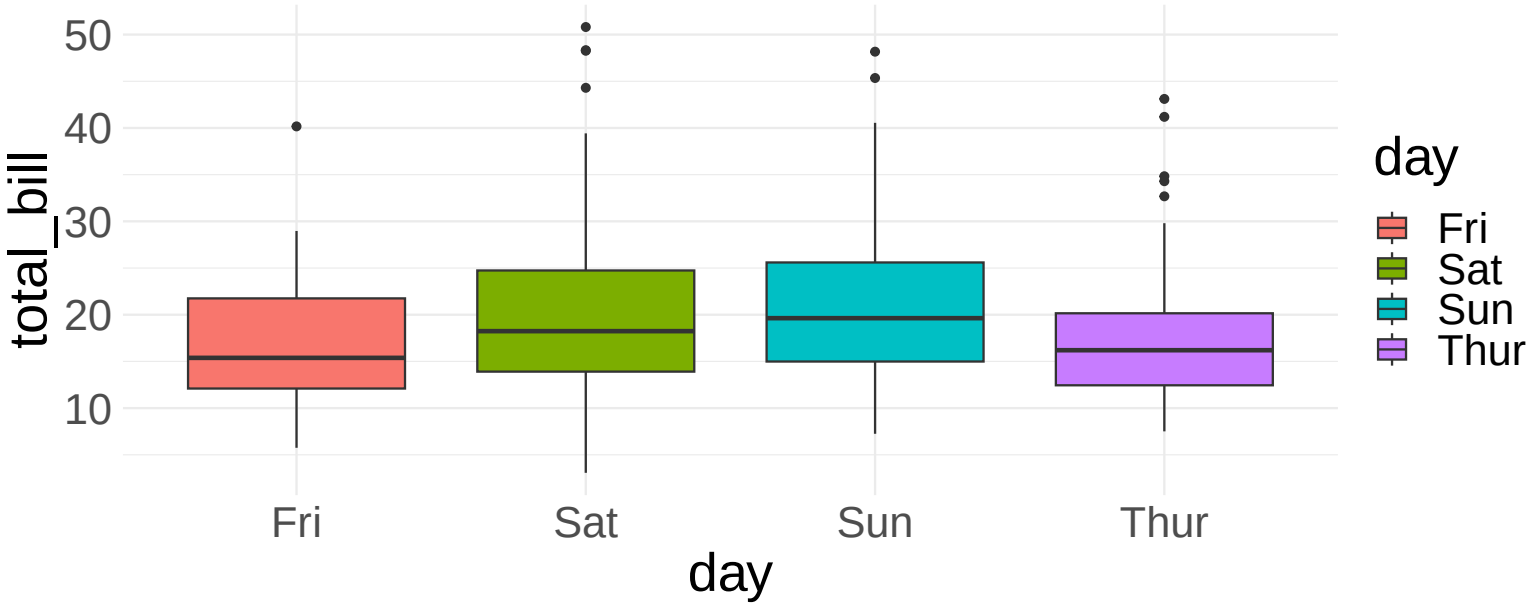
A regarder aussi :

- 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 des hypothèses faites sur la population en analysant l'échantillon

Outils

- Estimation : approximer des paramètres de la population
- Test : valider les hypothèses
- Modélisation : rechercher des liens entre variables