Chapitre 2 - Statistique Descriptive

Marouane Il Idrissi

il_idrissi.marouane@uqam.ca

STT 1000 - Automne 2025

Département de Mathématiques, Université du Québec à Montréal

Statistique descriptive

Définition *(Statistique descriptive)*.

La **statistique descriptive** est un **ensemble de méthodes** (représentations graphiques et calculs de caractéristiques numériques) permettant de **faire une synthèse statistique** à partir de **données**.

On s'intéresse à **un phénomène**
Le cyclisme à Montréal, la fission nucléaire, l'érosion des côtes...

On réalise des **expériences** et on collecte des données
Sondages, capteurs, mesures...

On **analyse ces données** pour **mieux comprendre le phénomène**
☞ **C'est là qu'entre en jeu la statistique descriptive**

Illustration en utilisant `R` et l'environnement de développement `RStudio`

Plan du chapitre

1. Analyse de données
2. Mesures de position
3. Mesures de dispersion
4. Représentations graphiques univariées
5. Représenter deux variables

Sommaire

Analyse de données - Variable, population, individu

Un **individu** est l'objet décrit par une donnée

Une personne, un assuré, un arbre...

Une **variable** est une **caractéristique** d'un individu

L'âge, le revenu, la côte de crédit, la circonférence...

La **population** est **l'ensemble des individus** que l'on souhaite étudier

Les habitants du Canada, les clients d'une compagnie, les arbres du parc national de Forillon...

Un **échantillon observé** est une **sous-partie de la population** dont les **variables sont mesurées**

Dans ce cours:

- On notera n la **taille de l'échantillon**
- On supposera que l'**échantillon est aléatoire**
  - ☞ Échantillon tiré au hasard dans la population, chaque individu a la même chance d'être tiré

Un **jeu de données**, c'est **l'échantillon observé** sous forme de **tableau de données**

Individus en ligne et variables en colonne

Analyse de données - Types de variables

Les **variables catégorielles** (ou **qualitatives**) **partitionnent** les individus en plusieurs groupes

- **Variables qualitatives nominales** ★
  Pas d'ordre dans les modalités
  Par ex: Couleur de la voiture ✍
- **Variables qualitatives ordinales** ★
  Il y a un ordre dans les modalités
  Par ex: Niveau de satisfaction d'un client ✍

Les **variables quantitatives** prennent des **valeurs numériques**
On peut les additionner, les multiplier...

- **Variables quantitatives discrètes** ★
  Prend des valeurs dans un ensemble dénombrable
  Par ex: Nombre d'enfant dans une famille ✍
- **Variables quantitatives continues** ★
  Prend des valeurs dans un intervalle
  Par ex: Diamètre (cm) d'un arbre ✍

Pour illustrer ce chapitre, nous allons étudier le **jeu de données**:
Données Ouverte de la Ville de Montréal - Inventaire des arbres publics de la ville

☞ **Qu'est-ce qu'un individu?** ✍

☞ **Quelle est la population?** ✍

Les variables sont:

- `arrondissement`: Arrondissement de la ville de Montréal
- `lieu`: Type de lieu (parc, parterre, trottoir)
- `type`: Type d'arbre
- `dhp`: Diamètre à hauteur de poitrine de l'arbre (cm)
- `remarquable`: Arbre remarquable ou non
- `age`: Âge (années) depuis la plantation
- `maturité`: Nombre de décennies depuis la plantation

☞ **Quel est le type de chacune de ces variables?** ✍

Voyons ce que ça donne sur le logiciel R!

Sommaire

Mesures de position - Moyenne

Une **mesure de position** indique la **tendance centrale** d'un ensemble d'observations

La **moyenne** $\overline{x}$ d'un **ensemble d'observations** $x_1, \ldots, x_n$:

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

Example *(Performance d'une voiture)*. Nombre de litres par 100km pour un échantillon de 13 voitures:
$9, 8, 9, 12, 14, 10, 3, 15, 18, 11, 14, 9, 9$
☞**Quelle est la consommation moyenne des voitures dans l'échantillon?**

Définition *(Centrer une variable)*. ★

Soit $x_1, \ldots, x_n$ un échantillon observé. **Centrer** cet échantillon revient à **retirer la moyenne à chaque observation**. C'est l'échantillon $\breve{x}_1, \ldots \breve{x}_n$, où, pour $i = 1, \ldots, n$:

$$\breve{x}_i = x_i - \overline{x} = x_i - \frac{1}{n}\sum_{i=1}^{n} x_i$$

☞ **Quelle est la moyenne de l'échantillon $\breve{x}_1, \ldots, \breve{x}_n$?** ✍

Variables qualitatives - Fréquence et proportion

Pour les **variable qualitative**, on peut calculer les **fréquences d'apparition des modalités**

La **fréquence** d'une modalité, c'est **compter combien de fois elle apparaît**

Pour un **échantillon observé** d'une **variable qualitative à m modalités** $M_1, \ldots, M_m$, la **fréquence** de la modalité M_j se calcule par:

$$\mathrm{Freq}_{M_j} = \sum_{i=1}^{n} \mathbb{1}_{M_j}(x_i), \quad \text{où} \quad \mathbb{1}_{M_j}(x_i) = \begin{cases} 1, & \text{si } x_i = M_j \\ 0, & \text{si } x_i \neq M_j \end{cases}$$

La **proportion** d'une modalité M_j est donnée par:

$$\mathrm{Prop}_{M_j} = \frac{1}{n}\mathrm{Freq}_{M_j} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{M_j}(x_i)$$

☞ La **proportion d'une modalité** M_j, c'est la **moyenne empirique** de l'échantillon observé $\mathbb{1}_{M_j}(x_1), \ldots, \mathbb{1}_{M_j}(x_n)$ ★

Mesures de position - Médiane

Pour calculer la **médiane** $q_{0.5}$ d'un **échantillon observé** $x_1, \dots, x_n$:

1. On **range par ordre croissant** les n observations $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
   ☞ $x_{(1)}$ est **l'observation la plus petite** et $x_{(n)}$ **la plus grande**

2. La **médiane** $q_{0.5}$ est égale à:
   - $x_{\left(\frac{n+1}{2}\right)}$ si n **est impair**
   - $\frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$ si n **est pair**

Interprétation: 50% des individus de l'échantillon on une **valeur inférieure ou égale à la médiane** $q_{0.5}$ et 50% ont une valeur supérieure ou égale à la médiane.

☞ **Échantillon rangé par ordre croissant:** $3, 8, 9, 9, 9, 9, 10, 11, 12, 14, 14, 15, 18$
Quelle est la consommation médiane des voitures dans l'échantillon? ✍

☞ Qu'arrive-t-il à la médiane dans l'exemple précédent si on ajoute une nouvelle observation égale à 6? ✍

☞ **La médiane est une mesure de position plus robuste que la moyenne** ★✍

Mesures de position - Quantile

La **médiane** est un **cas particulier d'un quantile**

Pour $\alpha \in (0,1)$, le **quantile d'ordre** α, noté q_α, **de l'échantillon observé** est l'observation telle que $(\alpha \times 100)\%$ des observations **sont en dessous**

A partir d'un échantillon observé **rangé par ordre croissant** $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, on a que ★

$$q_\alpha = (1-\gamma)\, x_{(k)} + \gamma\, x_{(k+1)}$$

où $k = \lfloor (n+1)\alpha \rfloor$ et $\gamma = (n+1)\alpha - k$

Il y a d'autres méthodes pour calculer un quantile, **dans ce cours on utilisera celle-ci**

☞ Pour $\alpha = 0.5$, on retrouve **exactement la définition de la médiane** ✍

☞ Pour $\alpha = 1/4$, on parle de **premier Quartile** (souvent noté Q_1)

☞ Pour $\alpha = 3/4$, on parle de **troisième Quartile** (souvent noté Q_3)

☞ **Échantillon observé**: $3, 8, 9, 9, 9, 9, 10, 11, 12, 14, 14, 15, 18$

Quelle est la valeur de Q_1 et Q_3? ✍

Mesures de position - Moyenne tronquée

La **moyenne tronquée** à un niveau α (en général 0.1 ou 0.2) est **la moyenne en omettant les valeurs les** $(\alpha \times 100)\%$ **les plus petites et les plus grandes**:

$$\overline{x}_\alpha = \frac{x_{(\lfloor n\alpha+1 \rfloor)} + \cdots + x_{(n-\lfloor n\alpha \rfloor)}}{n - 2\lfloor n\alpha \rfloor}$$

☞ **Plus robuste** que la **moyenne de l'échantillon**
On ne prend pas en compte les très petites et très grandes valeurs

☞ **Échantillon observé**: $3, 8, 9, 9, 9, 9, 10, 11, 12, 14, 14, 15, 18$
Quelle est la valeur de la moyenne tronquée à $\alpha = 0.1$? ✍

Mesures de position - Mode

Le **mode** est l'observation qui **revient le plus souvent dans l'échantillon**

☞ **Échantillon observé**: $3, 8, 9, 9, 9, 9, 10, 11, 12, 14, 14, 15, 18$
Quelle est la valeur du mode? ✍

Pour des des **variables aléatoires**:

- **Discrètes:** Le mode est la **valeur où la fonction de masse est la plus grande**
- **Continues:** Le mode est la **valeur où la densité est la plus grande**

Sommaire

Mesures de dispersion - Variance

Une **mesure de dispersion** est un indicateur de **l'étendue des valeurs prises par une variable**

La **variance** s^2 d'un **échantillon observé** $x_1, \ldots, x_n$:

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

et **l'écart-type** d'un **échantillon observé** est $s = \sqrt{s^2}$.

D'où vient ce $n-1$? Réponse au Chapitre 3!

Définition *(Centrer-réduire une variable)*. ★

Soit $x_1, \ldots, x_n$ un échantillon observé. **Centrer-réduire** cet échantillon c'est **retirer la moyenne, puis diviser chaque observation par l'écart-type**. C'est l'échantillon $z_1, \ldots z_n$, où, pour $i = 1, \ldots, n$:

$$z_i = \frac{x_i - \overline{x}}{s} = \frac{x_i - \frac{1}{n}\sum_{i=1}^{n} x_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

☞ **Quelle est la moyenne et la variance de l'échantillon** $z_1, \ldots, z_n$**?** ✍

Mesures de dispersion - Étendue et écart interquartile

L'**étendue** d'un **échantillon observé** $x_1, \ldots, x_n$ est la **différence entre la plus grande et la plus petite valeur**:

$$E = x_{(n)} - x_{(1)}$$

La **variance** et **l'étendue** ne sont pas robustes

Une très petite ou très grande valeur peut grandement les influencer

Pour **résoudre ce problème**, on peut utiliser **l'écart interquartile**

C'est la **différence entre le troisième et le premier quartile**:

$$I_{\mathrm{IQ}} = Q_3 - Q_1$$

☞ **Échantillon observé**: $3, 8, 9, 9, 9, 9, 10, 11, 12, 14, 14, 15, 18$

Quelle est la valeur de la variance, de l'étendue et de l'écart interquartile? ✍

Sommaire

Représentations graphiques

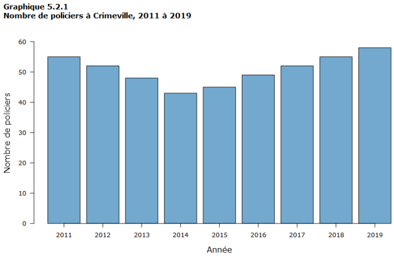
Jusqu'à présent, nous avons **résumé les informations contenues dans le jeu de données** à l'aide d'**indicateurs numériques**

Même si ces résumés sont **très précis**, il reste difficile de se faire une **idée globale de la manières dont sont distribuées les données**

Pour se faire une idée **plus générale**, on a recourt à des **représentations graphiques** pour représenter l'échantillon observé.

Ressource additionnelle pour vos révisions: Chapitre 5 de la Formation Statistique Canada

Un **diagramme à barre** permet de représenter les **fréquences ou les proportions** des modalité d'une **variable qualitative**



Source: Formation Statistique Canada

- **Que représente un individu dans ce jeu de données?**
- **Quelle est la variable qualitative que l'on étudie?**
- **Quelles sont ses modalités?**

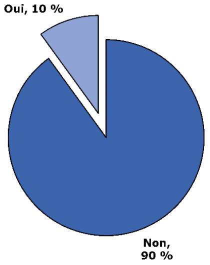
Les **diagramme à barre** permettent de **facilement comparer les occurrences des modalités dans un jeu de données**

Représenter une variable qualitative - Diagramme circulaire

Un **diagramme circulaire** remplit exactement la **même fonction qu'un diagramme à barre**

Comparer les **proportions des modalités** d'une **variable qualitative**

Graphique 5.4.1
Réponse des élèves et de la faculté à la question « Est-ce que les élèves de l'école Avenue devraient adopter l'uniforme? »

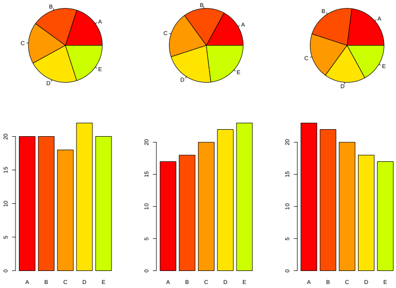


Source: Formation Statistique Canada

- ☞ **Que représente un individu dans ce jeu de données?**
- ☞ **Quelle est la variable qualitative que l'on étudie?**
- ☞ **Quelles sont ses modalités?**

Les **diagramme circulaire** permet une **comparaison rapide des modalités d'une variable qualitative dans un jeu de données**

Il est **difficile pour le cerveau humain** de **jauger les proportions représentées dans un cercle**



- **Pas plus de 5 modalités** (idéalement 2 ou 3)
- **Représenter toutes les modalités** (somme égale à 100)
- **Les écarts en % entre les modalités ne doivent pas être négligeables**
- **Ranger les modalités de la moins fréquente à la plus fréquente dans le sens horaire**
- **Ne pas comparer deux diagrammes circulaires**

Représenter une variable quantitative - Histogramme

Un **histogramme** permet de représenter **la distribution des données d'une variable quantitative**

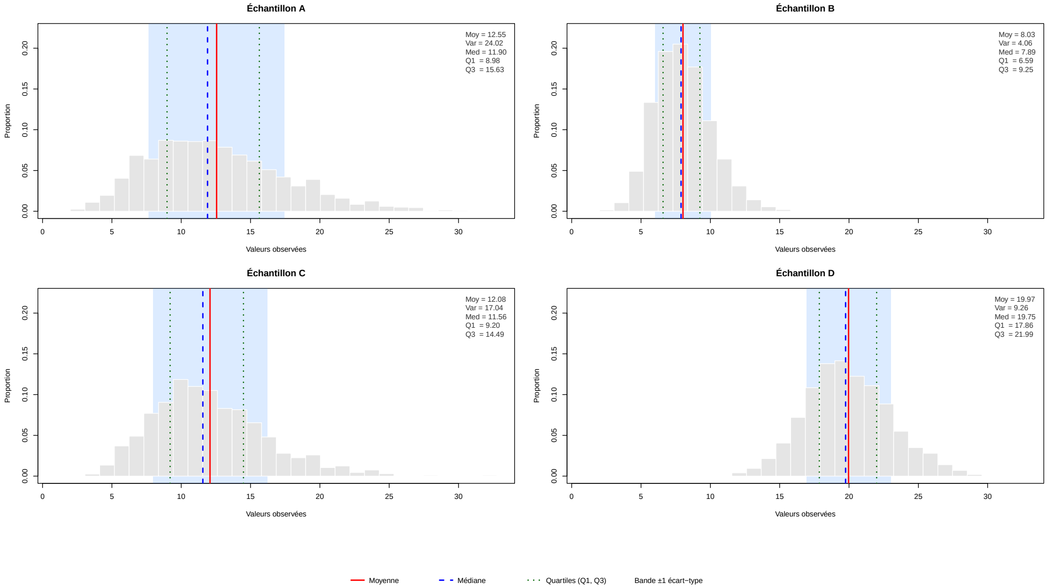
Pour **créer un histogramme**:

1. On **partitionne l'intervalle des valeurs de la variable**
   Généralement en **sous-intervalle de même taille**, mais ce n'est pas obligatoire
2. On compte **combien d'observations tombent dans chaque partition**
3. On trace des **rectangles pour chaque partition**, dont la largeur est **la largeur de la partition**, et la **hauteur est la proportion d'observations dans la partition**

On peut aussi choisir comme hauteur des rectangles la **fréquence** ou la **densité**

☞ **Plus on a de données observées, plus le partitionnement peut être fin (partitions petites)**

☞ **L'histogramme approche la densité de la variable aléatoire que l'on a observé**
Plus de détails dans les cours de statistique plus avancés


Échantillon A
Moy = 12.55
Var = 24.02
Med = 11.90
Q1 = 8.98
Q3 = 15.63
Échantillon B
Moy = 8.03
Var = 4.06
Med = 7.89
Q1 = 6.59
Q3 = 9.25
Échantillon C
Moy = 12.08
Var = 17.04
Med = 11.56
Q1 = 9.20
Q3 = 14.49
Échantillon D
Moy = 19.97
Var = 9.26
Med = 19.75
Q1 = 17.86
Q3 = 21.99
Proportion
Valeurs observées
Moyenne
Médiane
Quartiles (Q1, Q3)
Bande ±1 écart−type

Histogramme - Interprétation

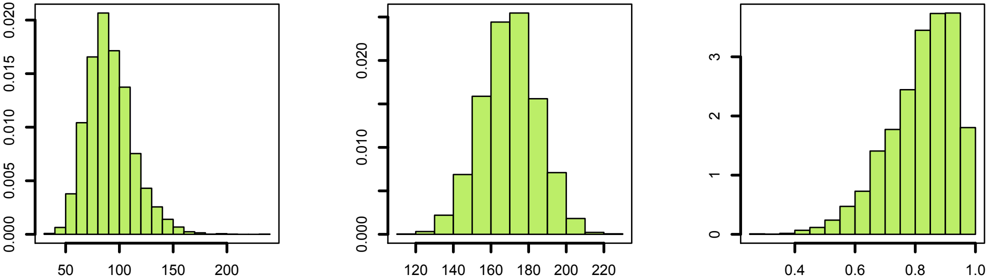
Les **histogrammes** permettent de décrire une distribution.

On dit que:

- Une distribution est **unimodale** si elle ne possède qu´un seul "pic" (mode)
- Une distribution est **symétrique** si les **queues de distribution ont la même longueur**
- Une distribution est **asymétrique à droite** si la **queue droite de la distribution est plus longue que celle de gauche**

  Une distribution est **asymétrique à gauche** si la **queue gauche de la distribution est plus longue que celle de droite**

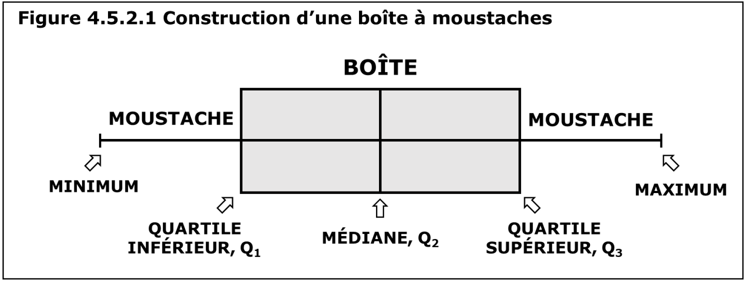
Histogramme - Symétrie



☞ **Commentez** ✍

Représenter une variable quantitative - Boîte à moustache

La **boîte à moustache** (boxplot) est une **représentation graphique de la dispersion d'un échantillon observé**



Source: Formation Statistique Canada

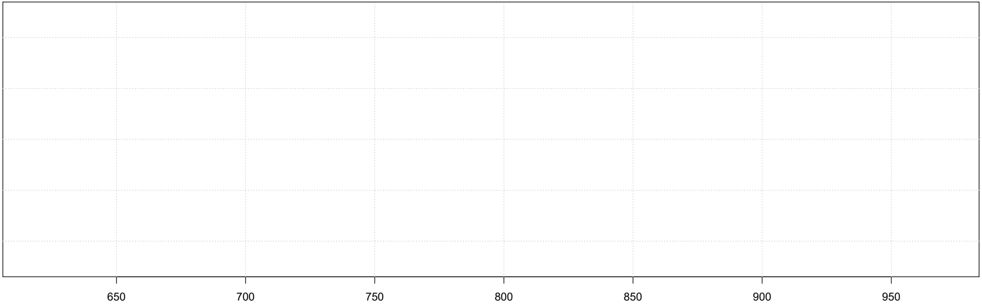
On peut **représenter les boîtes à moustache <u>verticalement</u> ou <u>horizontalement</u>**

Les points qui **sont plus loin qu'une fois et demi l'écart interquartile** sont considérés comme **potentiellement aberrante ou extrêmes** (outliers), et sont représentés **en dehors de la boîte**

Boîte à moustache - Exemple

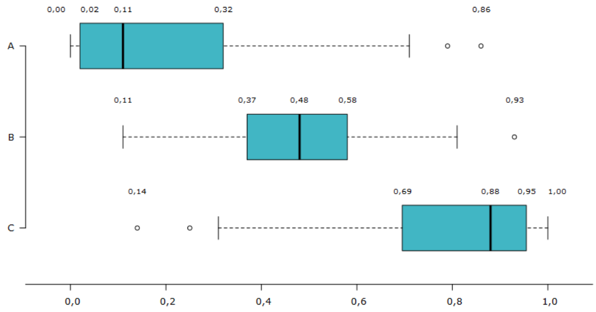
Mesure de la vitesse de la lumière par Albert Michelson (1879) en m/s (moins 299 000):

880, 880, 880, 860, 720, 720, 620, 860, 970, 950, 880, 910, 850, 870, 840, 840, 850, 840, 840, 840



Boîte à moustache - Interprétation

Graphique 4.5.2.1
Boîtes à moustaches et résumés en cinq nombres des distributions A, B et C



Source: Formation Statistique Canada

☞ Comparez ces 3 échantillons ✍

☞ Commentez leur symétrie ✍

<u>Boîte à moustache:</u>

Moins d'information qu'un histogramme, mais suffisamment pour pouvoir analyser la distribution de l'échantillon et propice à la comparaison entre variables ou échantillons

Sommaire

Représentation bivariées

Jusqu'à présent, on a analysé les variables **une par une**

Mais on peut être intéressés à exhiber **le lien entre deux variables**

Lorsque l'on met en relation **deux variables** on parle de **statistiques descriptives bivariées**

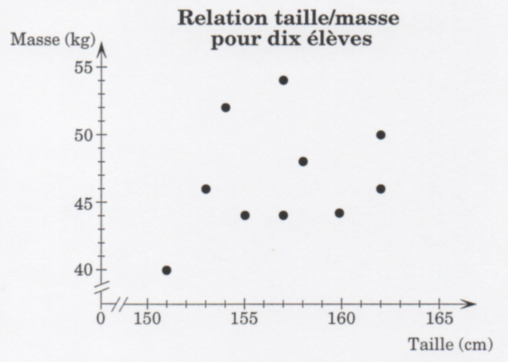
Représentation bivariées - Nuage de points

Un **nuage de point** permet de représenter la relation entre **deux variables quantitatives**

Cette représentation permet de **caractériser la nature du lien** entre deux variables

Linéaire, quadratique...

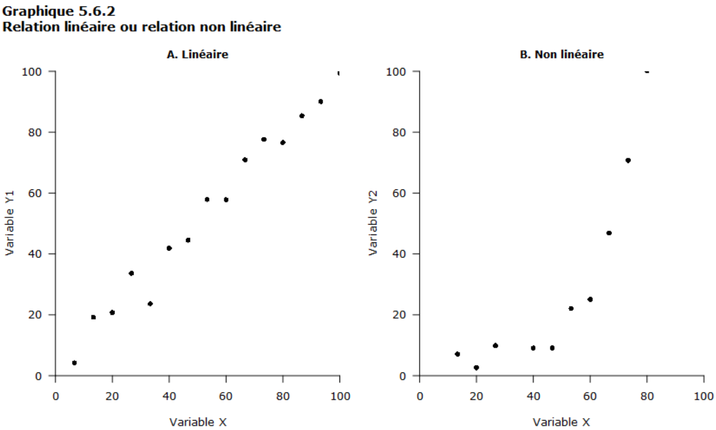
x	y
151	40
153	46
154	52
155	44
157	44
157	54
158	48
160	44
162	46
162	50



☞ On verra (si le temps) au Chapitre 7 comment **mesurer la force d'un lien linéaire**

Nuage de point - Lien linéaire ou non-linéaire

On dit qu'il y un **lien linéaire** entre deux variables si **on discerne une ligne droite entre les points du nuage**



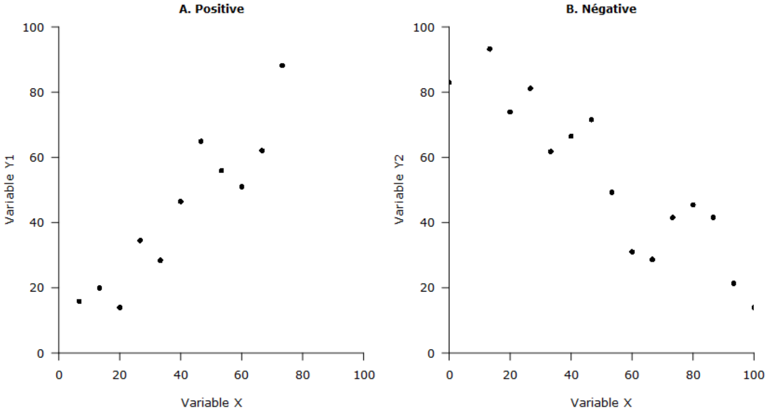
☞ Si on discerne **autre chose qu'une droite**, on dit que le **lien est non-linéaire**

Nuage de point - Lien linéaire positif ou négatif

Lorsque **les deux variables croissent ensemble**, on parle de **lien linéaire positif**

Si **une variable diminue lorsque l'autre augmente**, on parle de **lien linéaire négatif**

Graphique 5.6.3
Relation positive ou relation négative

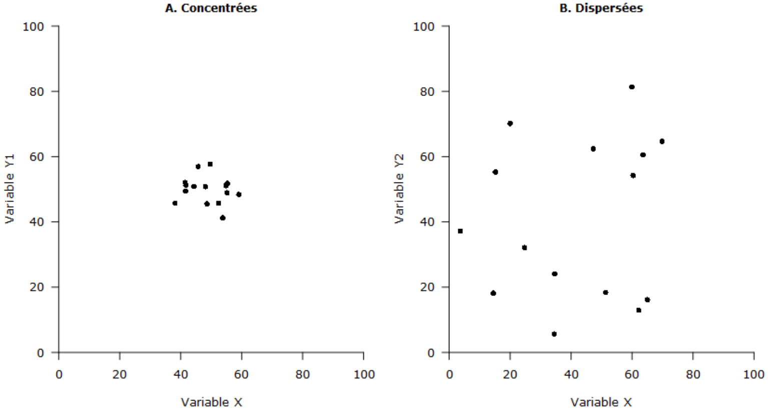


Nuage de point - Données concentrées et dispersées

Lorsque **les points sont rapprochés**, on dit que **les données sont concentrées**

Si **les points sont étalés**, on dit que **les données sont dispersées**

Graphique 5.6.4
Données concentrées ou données dispersées



Deux variables qualitatives - Tableau croisé

Lorsque l'on souhaite **comparer deux variables qualitatives**, on peut dresser un **tableau de fréquences croisées**

Pour une variable x à 3 modalités $\tilde{M}_1, \ldots \tilde{M}_3$, et une variable y à 4 modalités $M_1, \ldots, M_4$:

	M_1	M_2	M_3	M_4	Total
$\tilde{M}_1$					
$\tilde{M}_2$					
$\tilde{M}_3$					
Total					

Dans la **cellule relative** à la modalité $\tilde{M}_j$ et M_k, on indique

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\tilde{M}_j}(x_i)\mathbb{1}_{M_k}(y_i)$$

C'est-à-dire **le nombre d'individus ayant** $\tilde{M}_j$ **<u>et</u>** M_k **comme modalités aux variables** x **et** y

☞ **Que contiendront les cellules "Total"?** ✍

Deux variables qualitatives - Conditionnement

Pour étudier le lien entre une **variable qualitative et quantitative**, on peut **conditionner ou stratifier** l'échantillon observé

Cela revient à:

1. Créer un **nouvel échantillon conditionnel/stratifié par modalité observé**
   Sexe, profession...
2. Calculer les indicateurs ou représentation que l'on a vu jusqu'à présent **pour chaque nouvel échantillon conditionnel/stratifié**
3. Analyser les différences **entre les modalités**