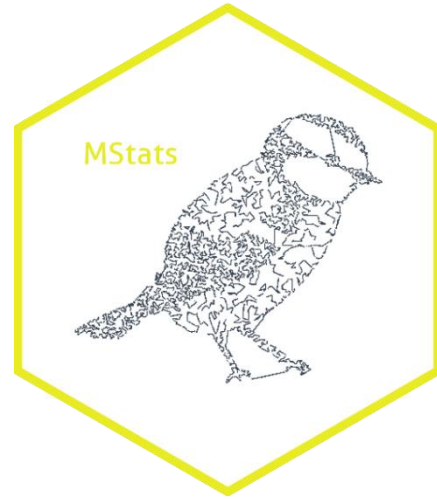
MQDS 2023
Measurement - Quality - Data Science



Data Analysis

Marie VAUGOYEAU
MStats
France

MQDS 2023
Measurement - Quality - Data Science

Avant de commencer

Marie VAUGOYEAU

Langage R et statistiques
Initiation à l'analyse de données

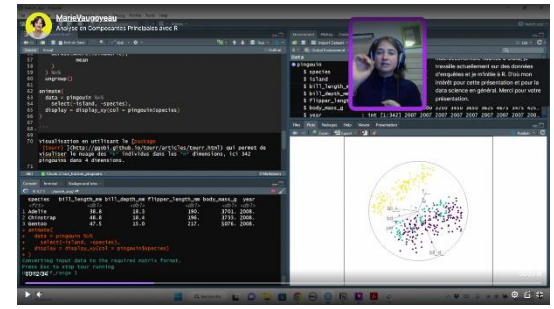
Auteur(s) : Marie VAUGOYEAU
Date de parution : 07/09/2022
Ref. ENI : RIRANADO
Collection : Ressources Informatiques
Expédié en 24h - en stock

Accompagnatrice indépendante à l'analyse de données et la formation au langage R

Dr en biologie évolutive et écologie comportementale

Partage de connaissances :

- Auteure de Langage R et Statistiques
- Rédactrice de la newsletter Aime les Stats et du blog Statistiques et R
- Réalisation de directs sur Twitch pour présenter des packages R et/ou des analyses statistiques



Tous ces liens et bien plus sur (https://linktr.ee/mstats)

MQDS 2023
Measurement - Quality - Data Science

Atelier d'analyse de données

Utiliser R pour valoriser ses données

Programme du jour

- Pourquoi R ?
- Création d'un projet R dans RStudio
- Production d'un fichier répétable en RMarkdown
- Réalisation d'une analyse descriptive :
  - Qu'est-ce que c'est ?
  - Pourquoi (s'embêter à) la faire ?
  - Comment faire ?
    - Représentation graphique global
    - Analyse descriptive univariée
    - Analyse descriptive bivariée puis multivariée

Pourquoi R ?

- Open-source (gratuit)
- Créé par et pour les statisticien.ne.s
- Langage spécialisé dans l'exploration et l'analyse de données
- Couramment utilisé dans les laboratoires et centres de recherches et les administrations

Création d'un projet R dans RStudio

- IDE le plus utilisé pour R
- Augmente la reproductibilité, facilite la transmission et améliore la stabilité
- Travail dans un « dossier » sans interaction avec les autres
- Chemin à la racine
- Structure minimale conseillée :
  - data_raw et data
  - img ou image
  - doc
- Possible de figer l'environnement
- Versionnage facile avec Git

RMarkdown

- Création de rapports automatisés
- Principe de ***WYSIWYW*** (*What you see is what you want*) contrairement à `Word` ou associé ***WYSIWYG*** (*What you see is what you get*)

- Syntaxe simple
- Plusieurs formats de sortie : `.pdf`, `.html`…
- Organisation :
  - En-tête facultative
  - Texte
  - Morceaux (chunk) de code en `R`, `Python`, `Java`…

Production d'un fichier répétable en RMarkdown

- Part belle au texte
- Automatisation et réutilisation facile des lignes de codes
- Enchainement d'images, graphiques, liens, tableaux, sortie de modèles facilité
- Syntaxe :
  - # pour les niveaux de titres
  - * pour l'italique
  - ** pour le gras
  - ``` pour le format code

Analyse descriptive : définition

- Trois types de statistiques :
  - Statistiques descriptives
  - Statistiques inférentielles ou probabilistes
  - Statistiques prédictives

- Les **statistiques** : Ensemble de méthodes qui ont pour objet la **collecte**, le **traitement** et l'**interprétation** de l'ensemble des **données d'observation** relatives à une **population statistique** (groupe d'individus ou d'unités) ou concernant un phénomène quelconque.

MQDS 2023
Measurement - Quality - Data Science

Le traitement et l'interprétation

- Le traitement et l'interprétation des données dépend de leurs natures, de leurs variations, de la quantité disponible, de la question posée…

➔ Il faut réfléchir pour analyser les données de **manière adaptée** et **interpréter correctement** les résultats

MQDS 2023
Measurement - Quality - Data Science

Pourquoi faire une analyse descriptive ?

- Première exploration des données
- Basées sur des graphiques et des calculs simples
- But :
  - Avoir un **premier aperçu** des données qui peut montrer des tendances.
  - **Caractériser les données**, ce qui est nécessaire pour choisir ensuite la manière de les analyser.
- Déroulement :
  - Aperçu graphique rapide
  - Statistiques **univariées** : Décrire les variables une par une
  - Statistiques **bivariées** : Explorer les variations d'une variable en fonction d'une autre

Types de données

- Variables qualitatives :
  - Dichotomiques (♀ ♂) ou pas
  - Nominale (couleur…) ou ordinale (taille de vêtement…)

- Variables quantitatives :
  - Discrètes (nombre de personnes,…)
  - Continues (masse, taille,…)

Description univariée

Variable qualitative

Description des variables qualitatives

- Type de modalités

- Nombre de modalités :
  - Infinies → variables simplificatrices :
    - Nombre de mots
    - Recherche de groupement (couleurs, sentiments…)
    - Etude des modalités les plus présentes…
  - Finies :
    - Tableau de contingence
    - Représentations graphiques

MQDS 2023
Measurement - Quality - Data Science

Distribution

Diagramme en barres

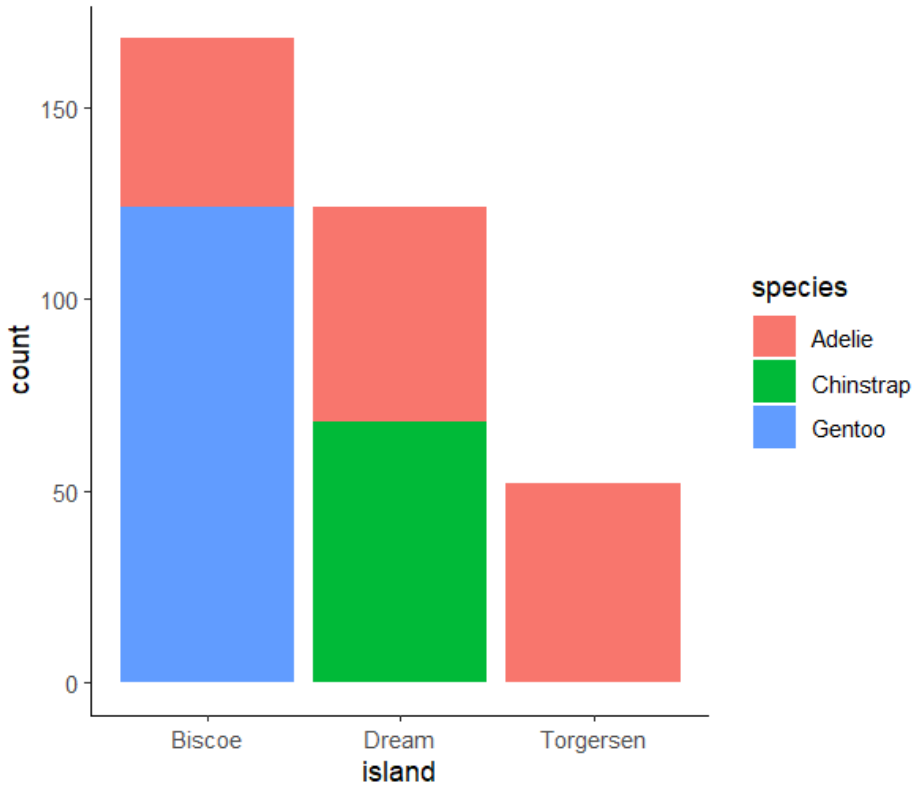
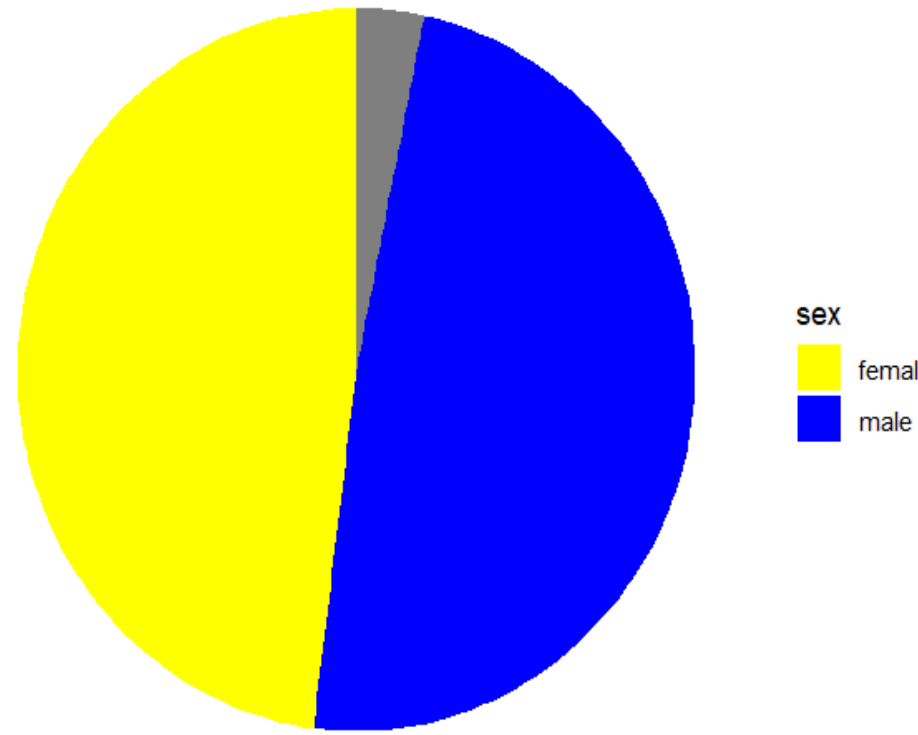


Diagramme circulaire

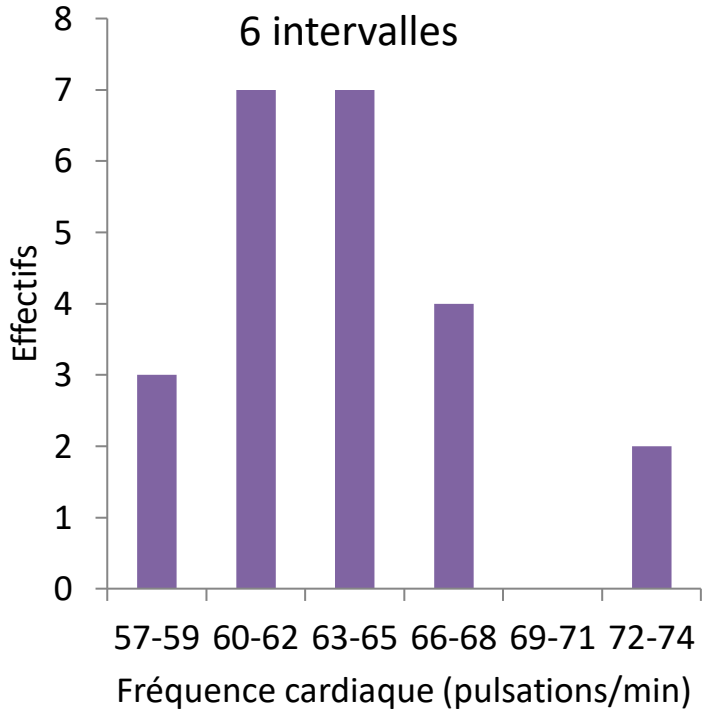


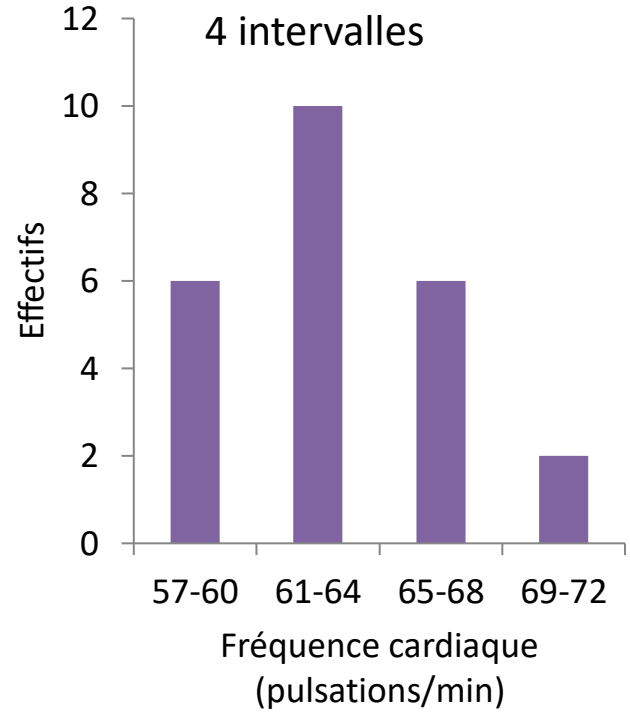
Description univariée

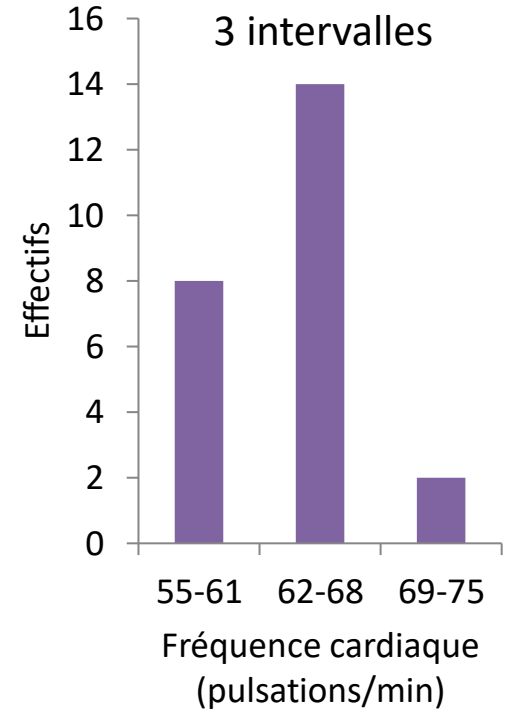
Variable quantitative

Histogrammes de la distribution

- Règle de Moore : Nombre d'intervalles environ égal à la racine carrée de l'effectif total







Description de la distribution

- **Le centre** : Valeur moyenne, valeur médiane
- **La dispersion** : Comment les valeurs s'écartent du centre (étendue, variance, écart-type)
- **La symétrie** : Répartition des données de part et d'autre du centre
- **Les points extrêmes** : Valeurs beaucoup plus faibles ou plus fortes que les autres

Décrire le centre

- Moyenne arithmétique : Somme des valeurs divisée par le nombre total de valeurs

$$\bar{x} = (x_1 + x_2 + x_3 + \ldots) / n$$

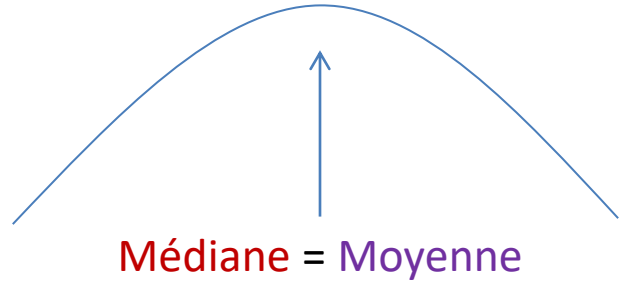
  - = centre des données si distribution symétrique
  - Sensible aux valeurs extrêmes

- Médiane : Valeur centrale si données triées par ordre
  - Une moitié des données > Médiane, l'autre moitié < Médiane
  - Valeurs extrêmes n'influencent pas la médiane ➔ plus robuste que la moyenne
  - Si nombre de données pair, médiane = valeur moyenne des 2 valeurs centrales
    - *Ex.* Données = 2.05 ; 3.56 ; 4.67 ; 6.90 ; 7.53 / Médiane : 4.67
    - *Ex.* Données = 2.05 ; 3.56 ; 4.67 ; 6.90 ; 7.53 ; 8.75 : Médiane : (4.67+6.90) /2 = 5.785

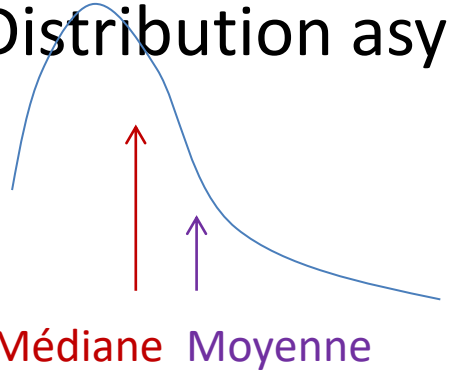
Moyenne, médiane et symétrie

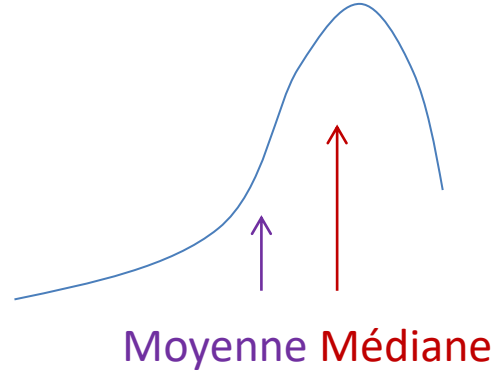
Position relative de la médiane et de la moyenne

- Distribution symétrique : Médiane = Moyenne



- Distribution asymétrique





MQDS 2023
Measurement - Quality - Data Science

La dispersion

- **L'étendue** : Différence entre la valeur maximale et la valeur minimale

$$e = x_{max} - x_{min}$$

  - Ne prend pas en compte l'ensemble des valeurs.

- **L'écart type** : Dépend de la déviation des valeurs par rapport à la moyenne ($x-\bar{x}$) et de l'effectif n de l'échantillon

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

  - Même unité que les données
  - Généralement, la grande majorité des données est à moins de 2 écarts types de la moyenne (entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$)

La dispersion (suite)

- **La variance** : Ecart type au carré

$$V = \sigma^2$$

  - On utilise généralement plus l'écart type que la variance

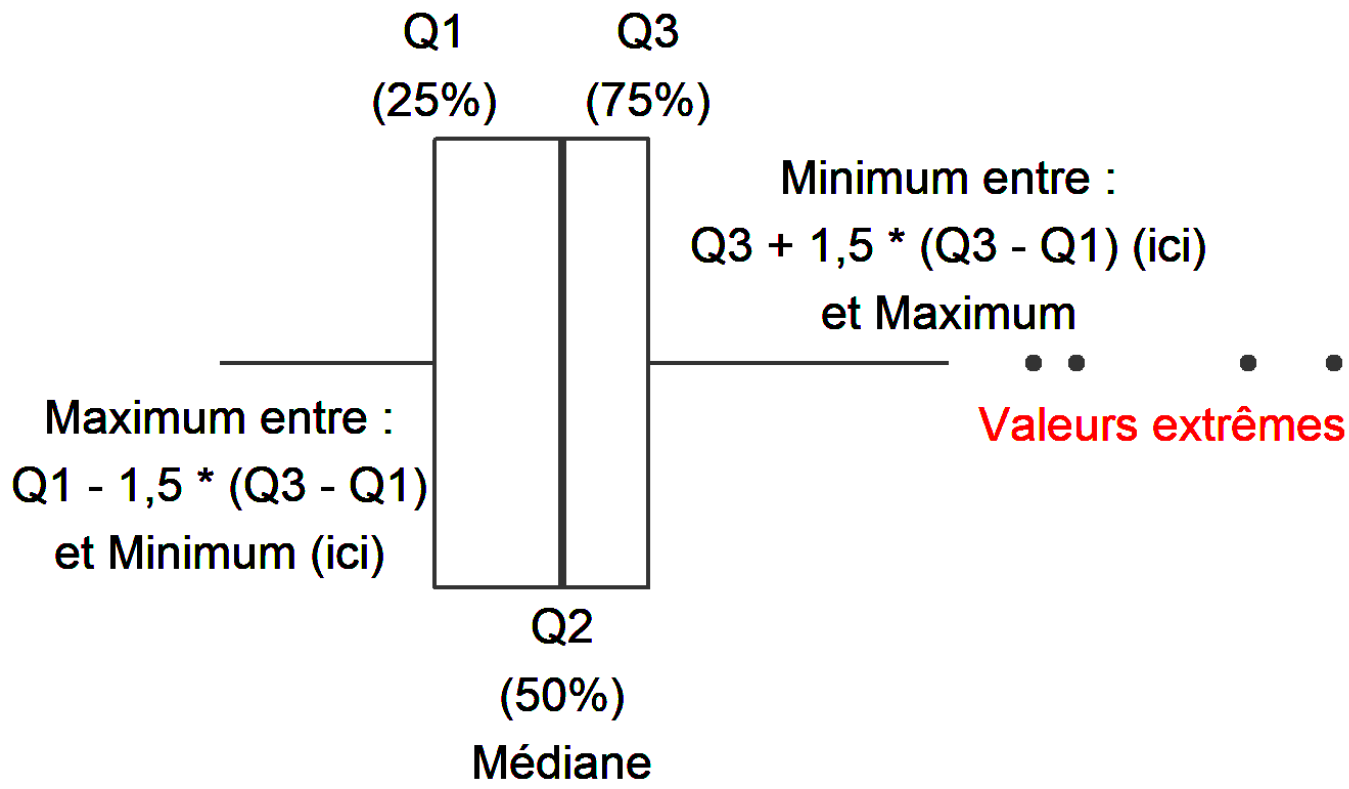
- **Les quartiles** : Sépare les données triées en 4 parties égales
  - Premier quartile (Q_1) sépare dans données triées les premiers 25% des 75% restants (médiane des données inférieures à la médiane)
  - Deuxième quartile (Q_2) sépare les premiers 50% des données triées des 50% restants (médiane)
  - Troisième quartile (Q_3) sépare les premiers 75% des données triées des 25% restants (médiane des données supérieures à la médiane)
  - Etendue interquartile (EIQ = Q_3 - Q_1) exprime la dispersion de la portion centrale des données

Ex. 81.6 91 92.5 92.5 99.8 110.3 118.8 130.7 150.6 156.0 157.9 159 163.3

Q_1 Q_2 Q_3

La dispersion : graphique de synthèse

Les boîtes à moustaches montrent si la distribution est symétrique ou non



Les points extrêmes

- Les valeurs extrêmes méritent qu'on s'y intéresse :
  - Possibilité d'erreur (de mesure, de frappe, …)

  ➔ Corriger ou retirer la valeur
  - Si valeurs confirmées

  ➔ Présente un intérêt (cas particulier…)

- Valeurs extrêmes < Q_1 - 1.5 EIQ *ou* > Q_3 + 1.5 EIQ

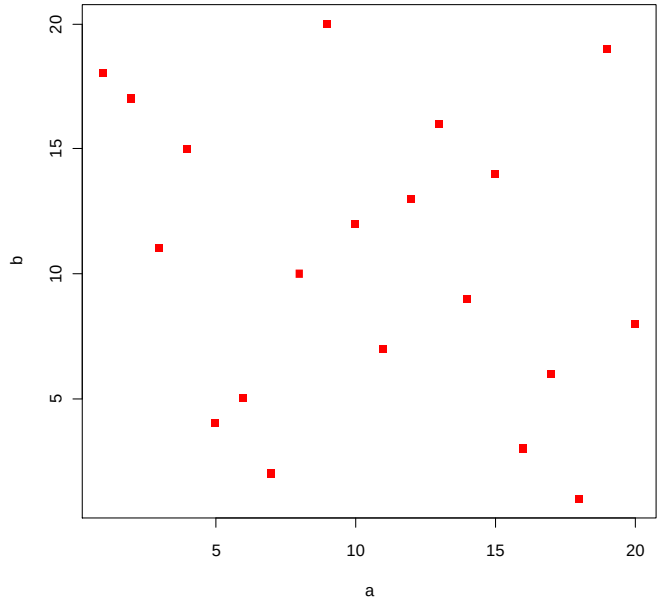
- Sur une boîte à moustaches, ces points sont représentés par des petits cercles à l'extérieur des moustaches

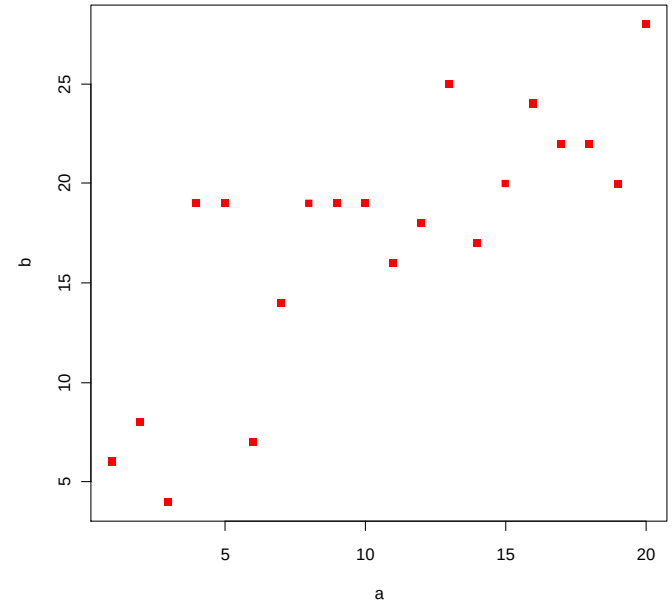
Description bivariée

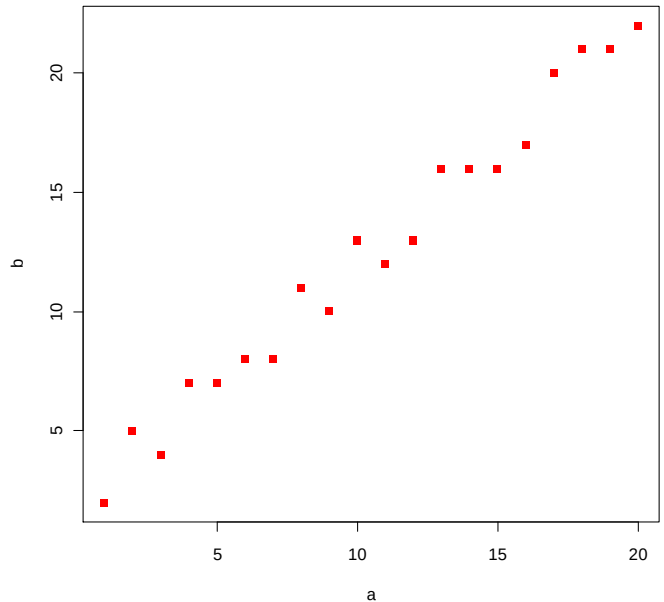
Deux variables quantitatives

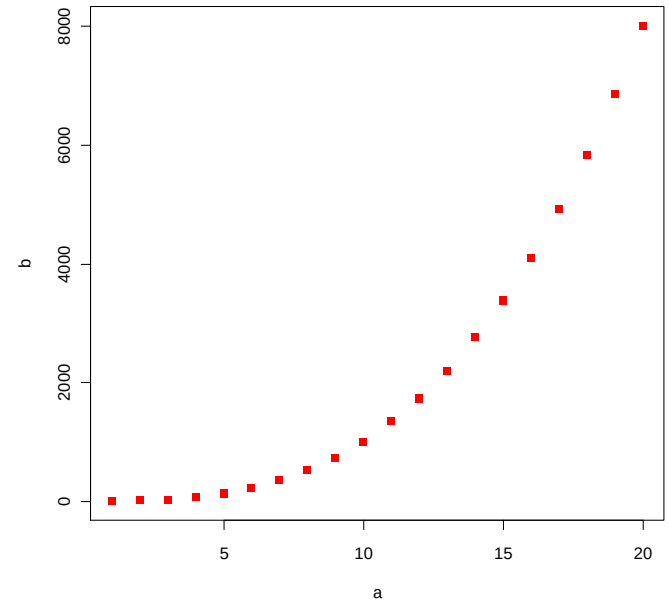
MQDS 2023
Measurement - Quality - Data Science

Le nuage de points



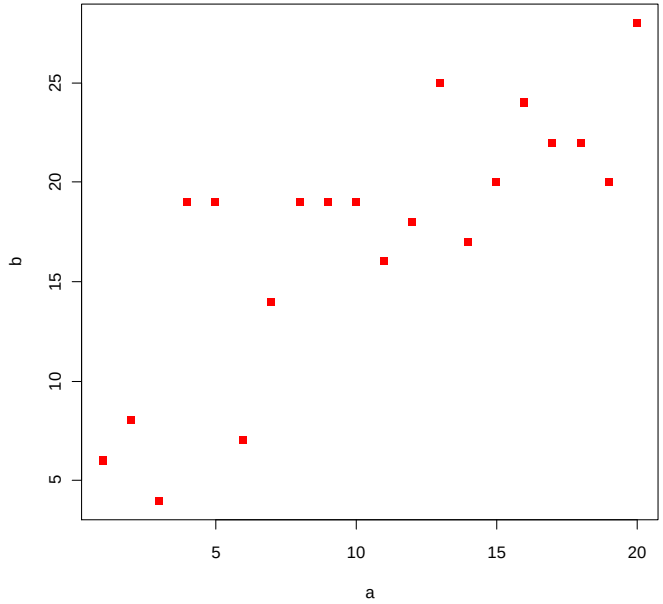




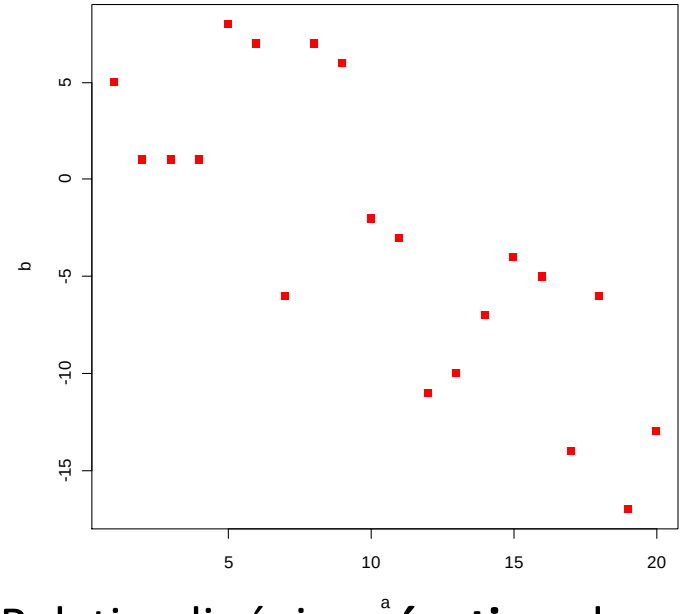


Le nuage de points

Une relation est linéaire lorsque le nuage de points paraît étiré le long d'une droite



Relation linéaire **positive** : Les deux variables évoluent dans le **même sens** (b tend à augmenter quand a augmente)

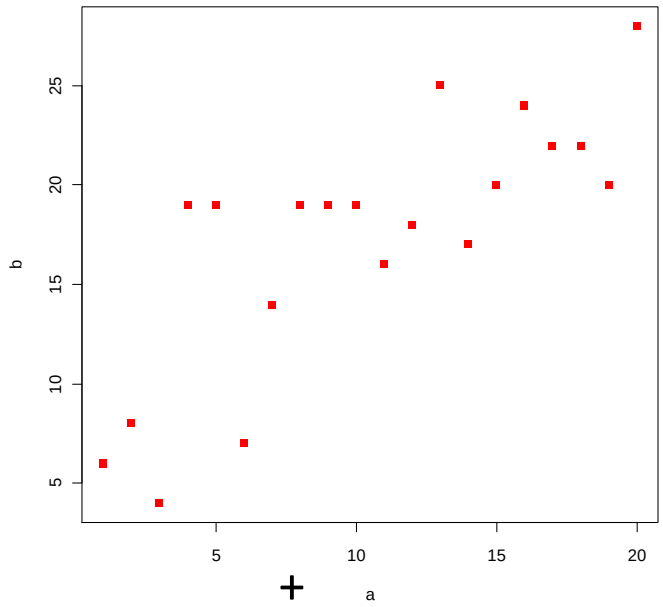


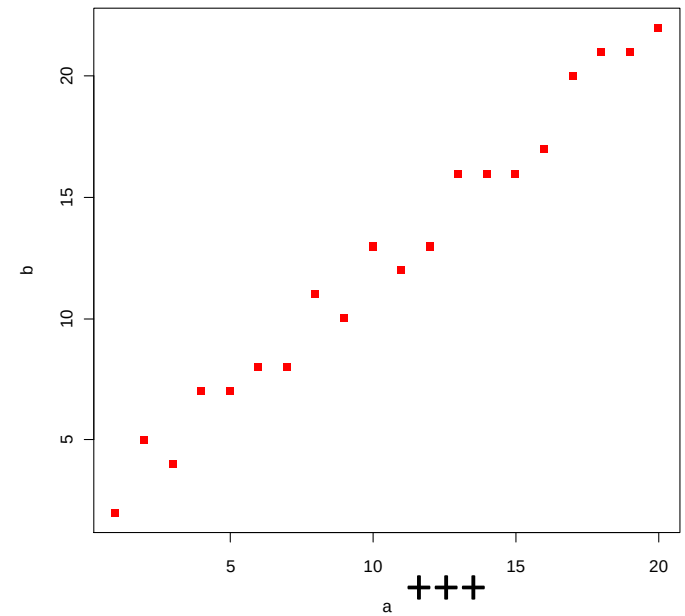
Relation linéaire **négative** : les deux variables évoluent en **sens contraire** (b tend à diminuer quand a augmente)

MQDS 2023
Measurement - Quality - Data Science

Le nuage de points

Plus les données s'organisent en droite, plus la relation entre les deux variables est forte





Le coefficient de corrélation linéaire (r)

- Définit l'intensité et le sens d'une relation linéaire entre deux variables quantitatives

- Toujours compris entre -1 et 1

- Le signe indique le sens de la relation et la valeur absolue indique l'intensité de la relation :
  - Proche de 1 : relation positive forte
  - Proche de 0 : relation très faible
  - Proche de -1 : relation négative forte

- Calcul :

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

$= \text{cov}(x,y) \times N$

$\sigma(x) \times \sqrt{N} =$

$= \sigma(y) \times \sqrt{N}$

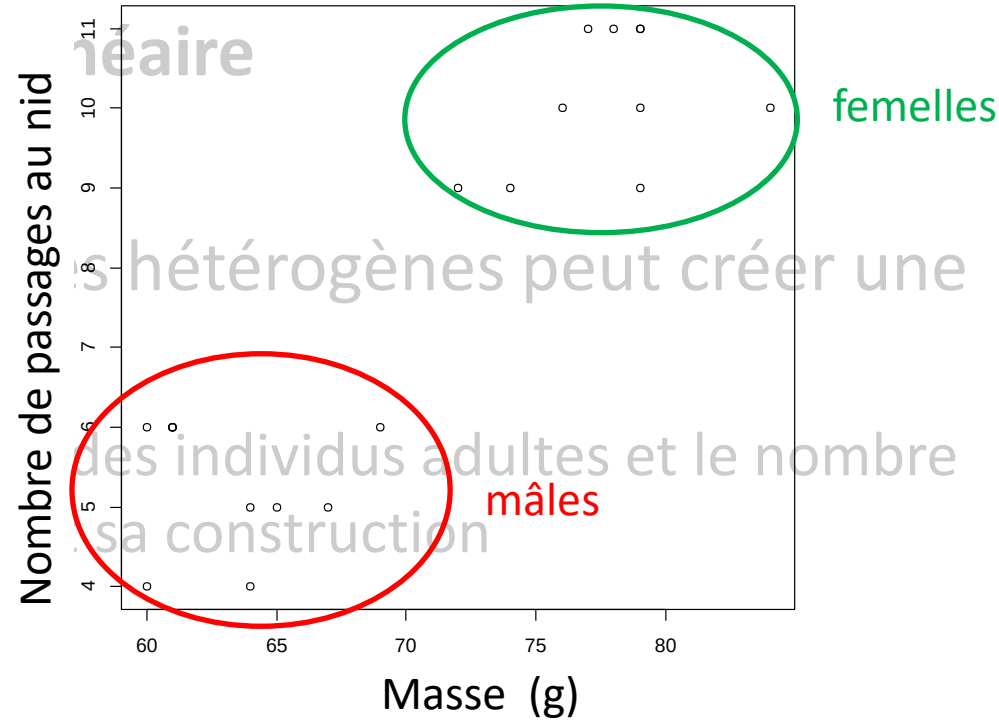
Le coefficient de corrélation linéaire (r)

- Quelques erreurs à ne pas commettre :
  - r fort ⇸ Relation de cause à effet entre les deux variable
  - r faible ⇸ pas de relation entre les deux variables Il peut y avoir une **relation non linéaire**

  - Une association de groupes hétérogènes peut créer une corrélation artificielle
    - *Ex.* Relation entre la masse des individus adultes et le nombre de passages au nid pendant sa construction

Le coefficient de corrélation linéaire (r)

- Quelques erreurs à ne pas commettre :
  - r fort ⇏ Relation de cause à effet entre les deux variable
  - r faible ⇏ pas de relation entre les deux variables Il peut y avoir une **relation non linéaire**
  - Une association de groupes hétérogènes peut créer une corrélation artificielle
    - *Ex.* Relation entre la masse des individus adultes et le nombre de passages au nid pendant sa construction



Le coefficient de détermination (r^2)

- Définie l'intensité de la relation entre les variables quelque soit son sens

- Varie entre 0 et 1

- Méthode : Pour savoir si la relation peut être considérée comme linéaire, on compare avec le seuil disponible dans une table. Le seuil dépend de l'effectif de l'échantillon.

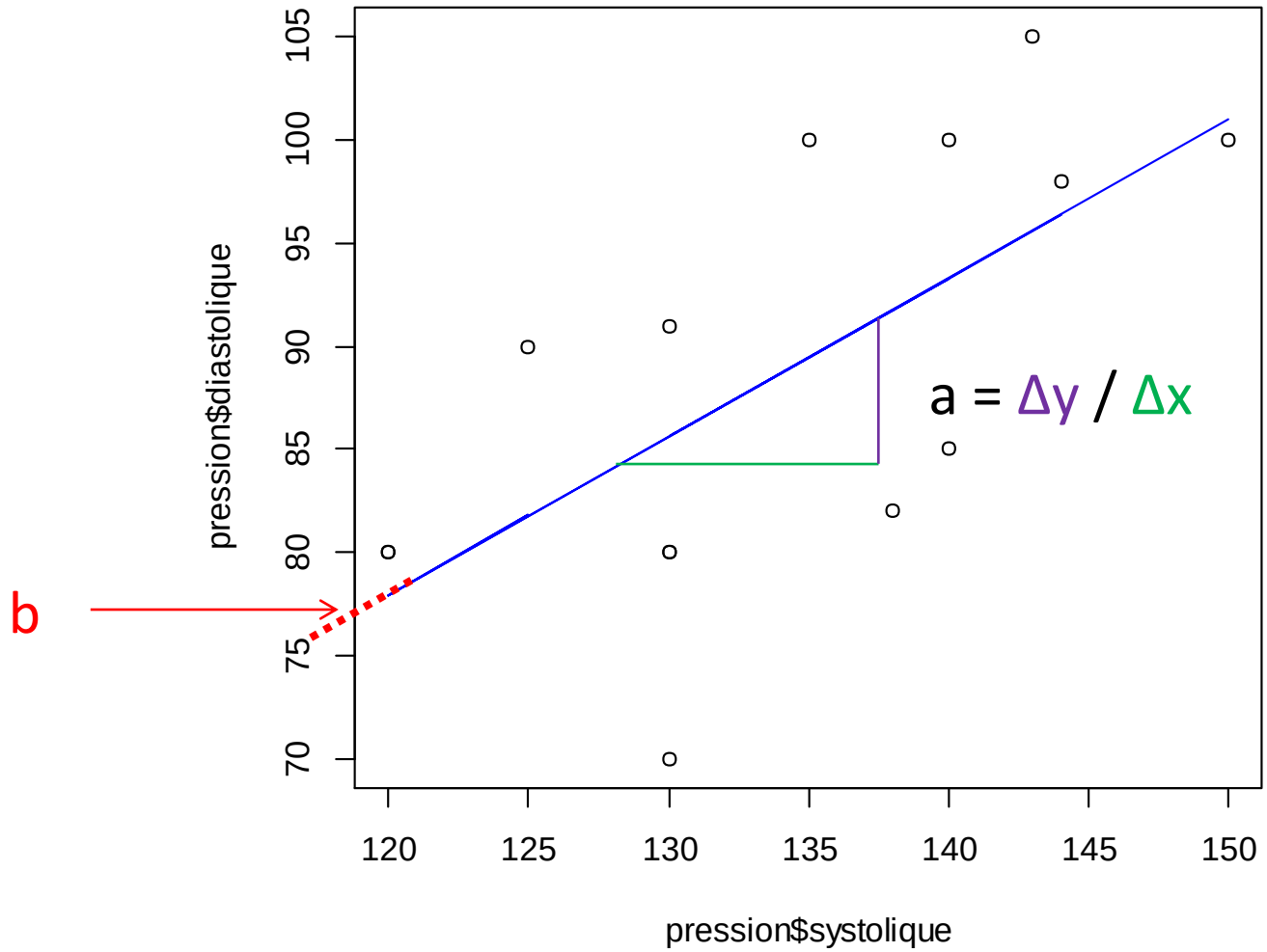
La droite de régression

- But de la régression : essayer de prévoir l'une des variables par rapport à l'autre

- Droite de régression : modélise la réponse moyenne de y en tout point d'abscisse x

- Son équation est de la forme y = ax + b, avec :
  - a : la pente de la droite de régression
  - b : l'ordonnée à l'origine,

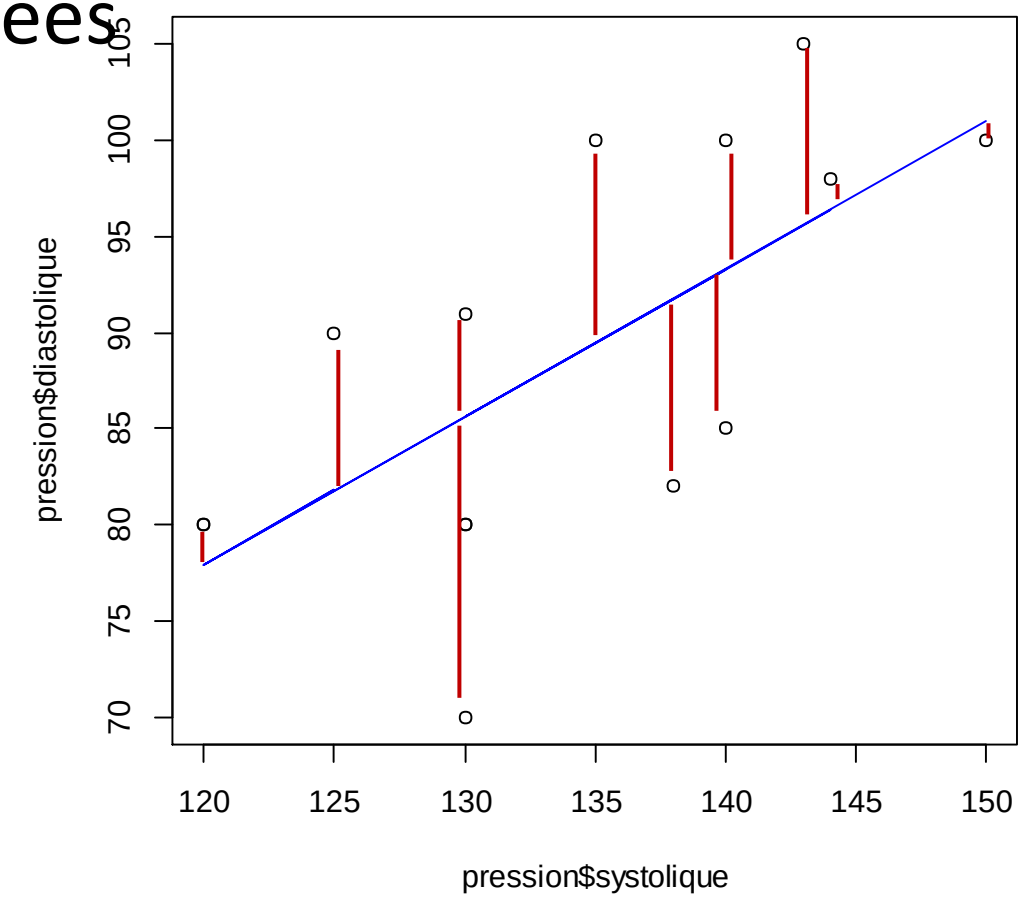
La droite de régression : y = ax + b



La droite de régression : les résidus

- Ecarts entre les valeurs prédites et les valeurs réellement mesurées

$$E = Y - (aX + b)$$



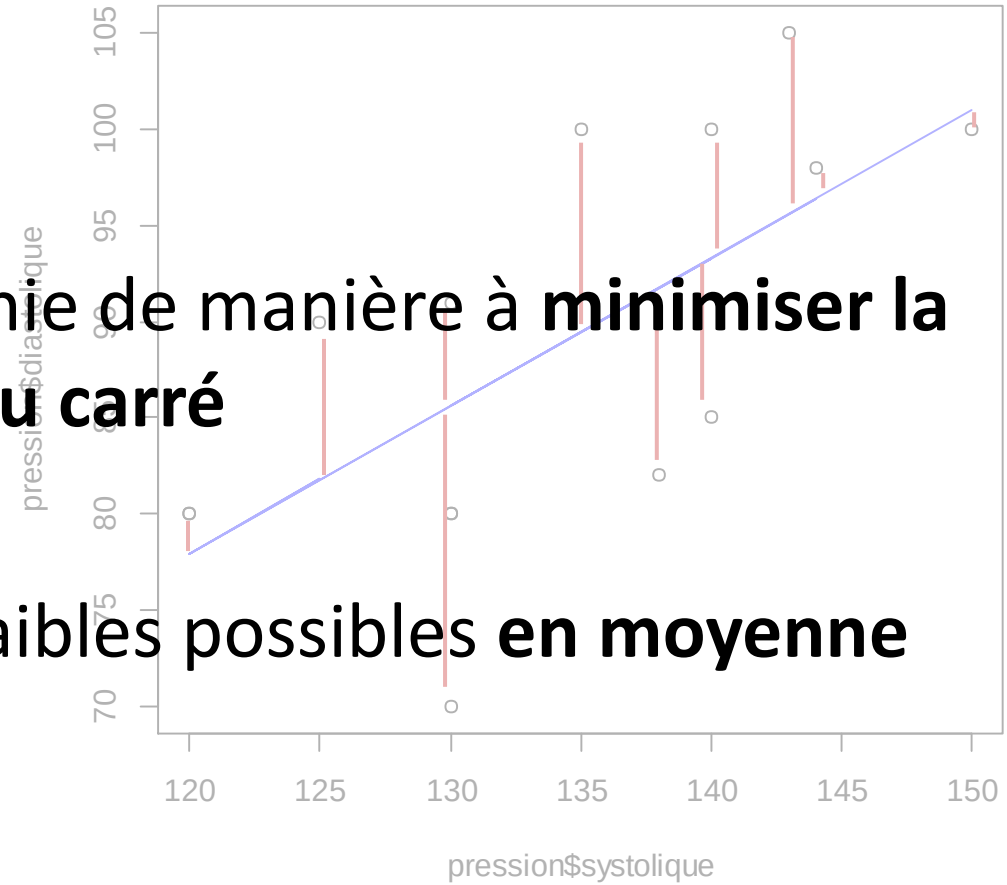
La droite de régression : les résidus

- Ecarts entre les valeurs prédites et les valeurs réellement mesurées

$E = Y - (aX + b)$

- Souvent la droite est définie de manière à **minimiser la somme des résidus mis au carré**

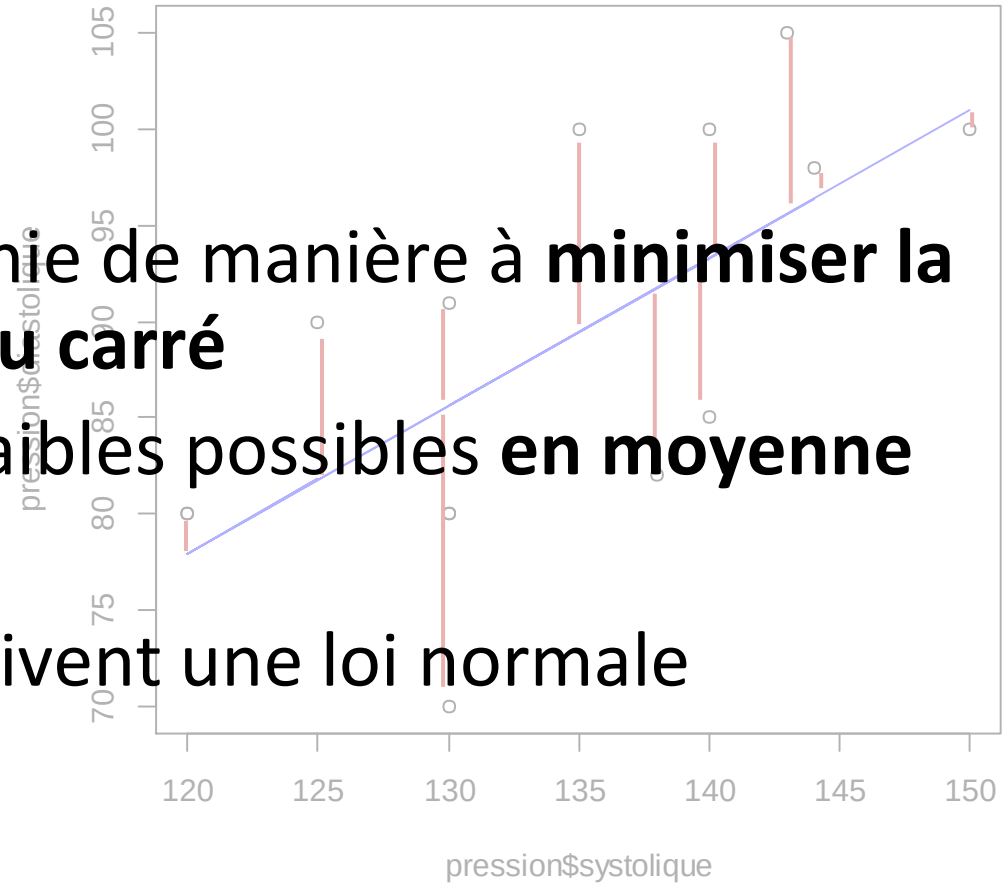
➔Avoir les écarts les plus faibles possibles **en moyenne**



La droite de régression : les résidus

- Ecarts entre les valeurs prédites et les valeurs réellement mesurées

$E = Y - (aX + b)$



- Souvent la droite est définie de manière à **minimiser la somme des résidus mis au carré**

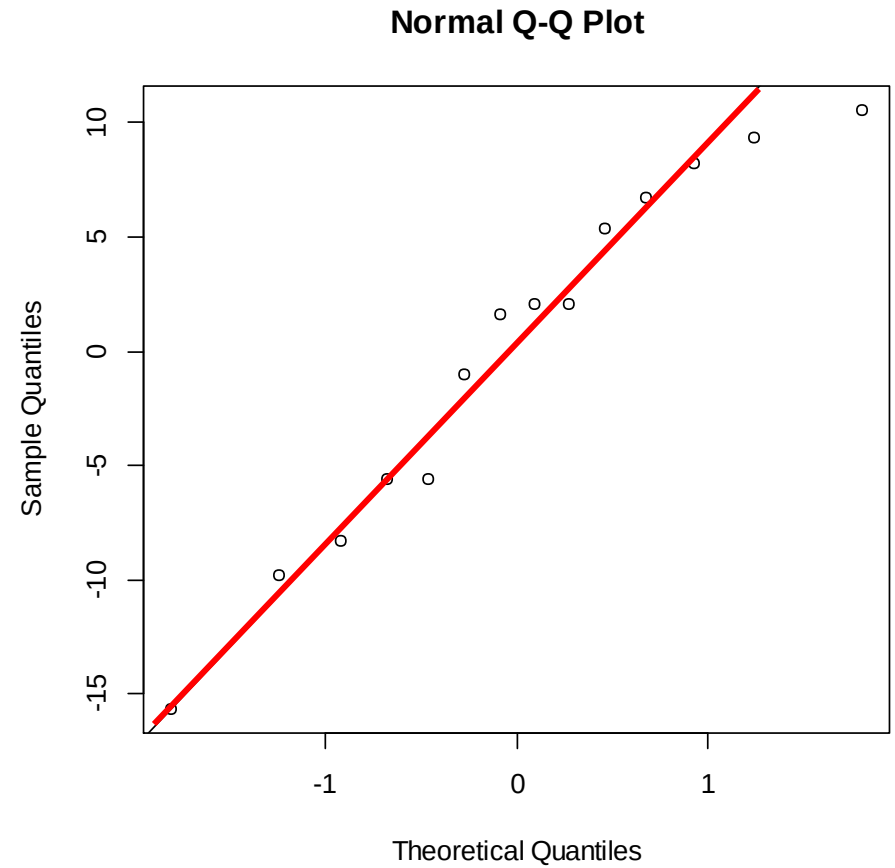
➔Avoir les écarts les plus faibles possibles **en moyenne**

- **Idéalement** les résidus suivent une loi normale

La droite de régression : les résidus

Le diagramme quantile-quatilde (Q-Q plot) montre l'adéquation de la distribution des résidus par rapport à une loi normale

Les points doivent-être le plus proche possible de la droite



Nouvelle interprétation de r^2

- Moyenne des prédictions = moyenne des valeurs mesurées

➔ En moyenne il n'y a pas d'erreur

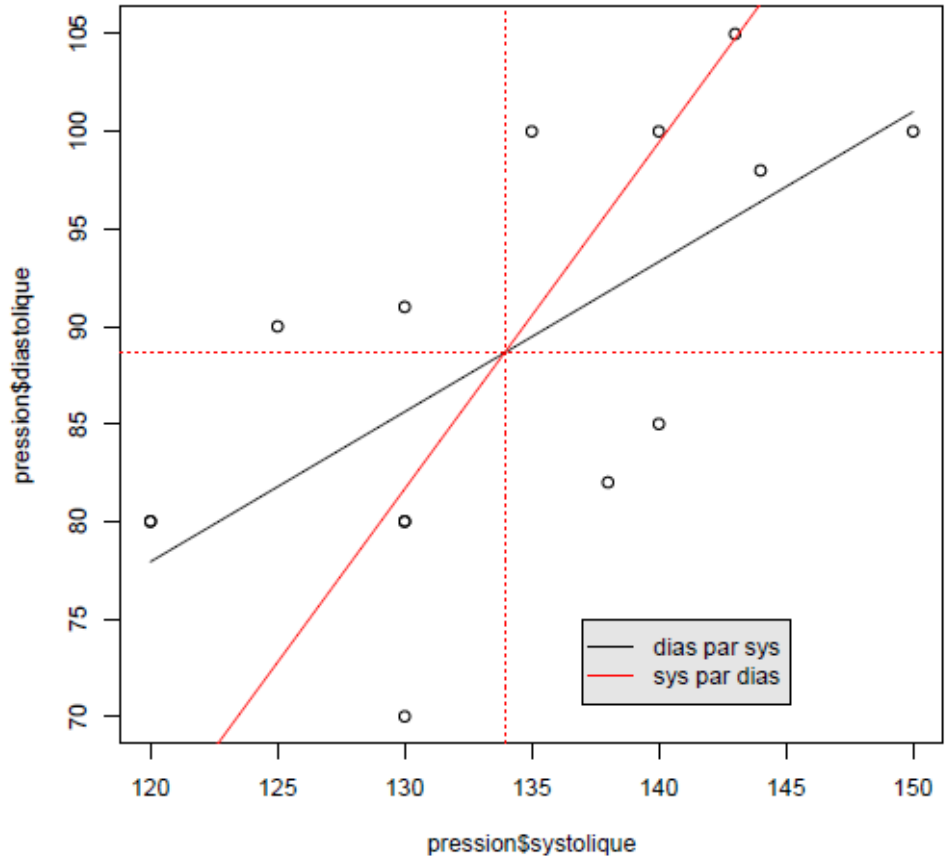
- Variance expliquée par la régression = variance des valeurs prédites
- La variance totale de la régression = variance expliquée + variance des erreurs (résidus)

- r^2 = Variance expliquée / Variance totale

Le coefficient de détermination est la **proportion de variance de la réponse Y pouvant être expliquée par la régression sur X**

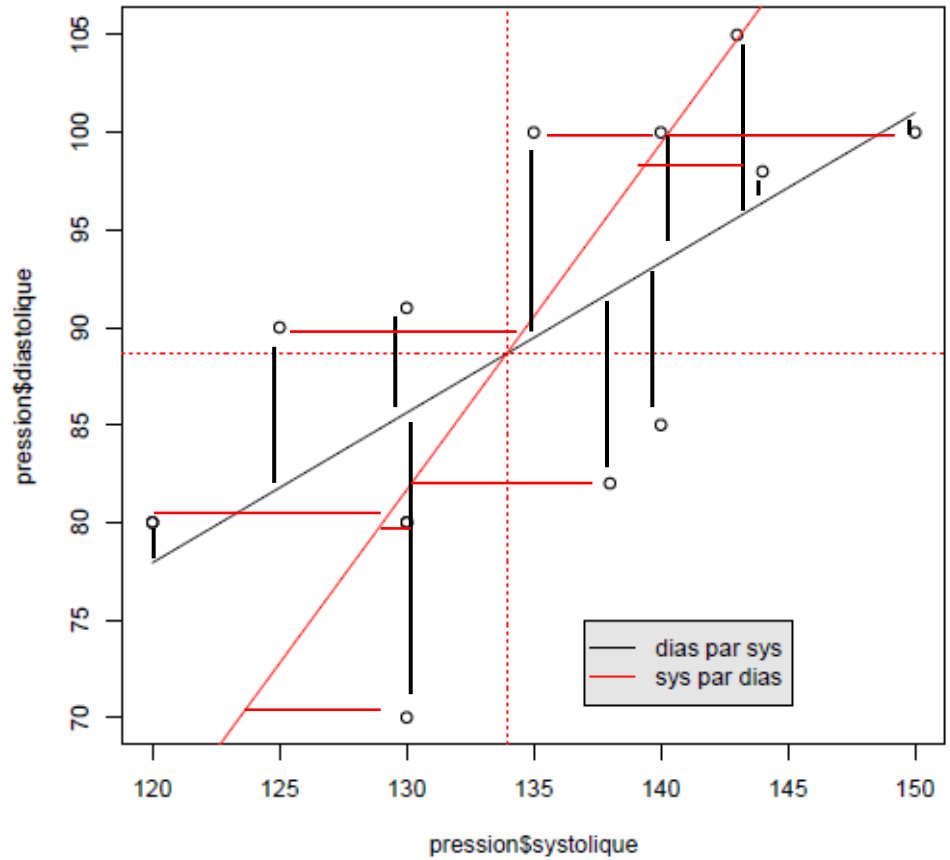
La régression : limite

- La régression de Y par X n'est pas la même chose que la régression de X par Y



La régression : limite

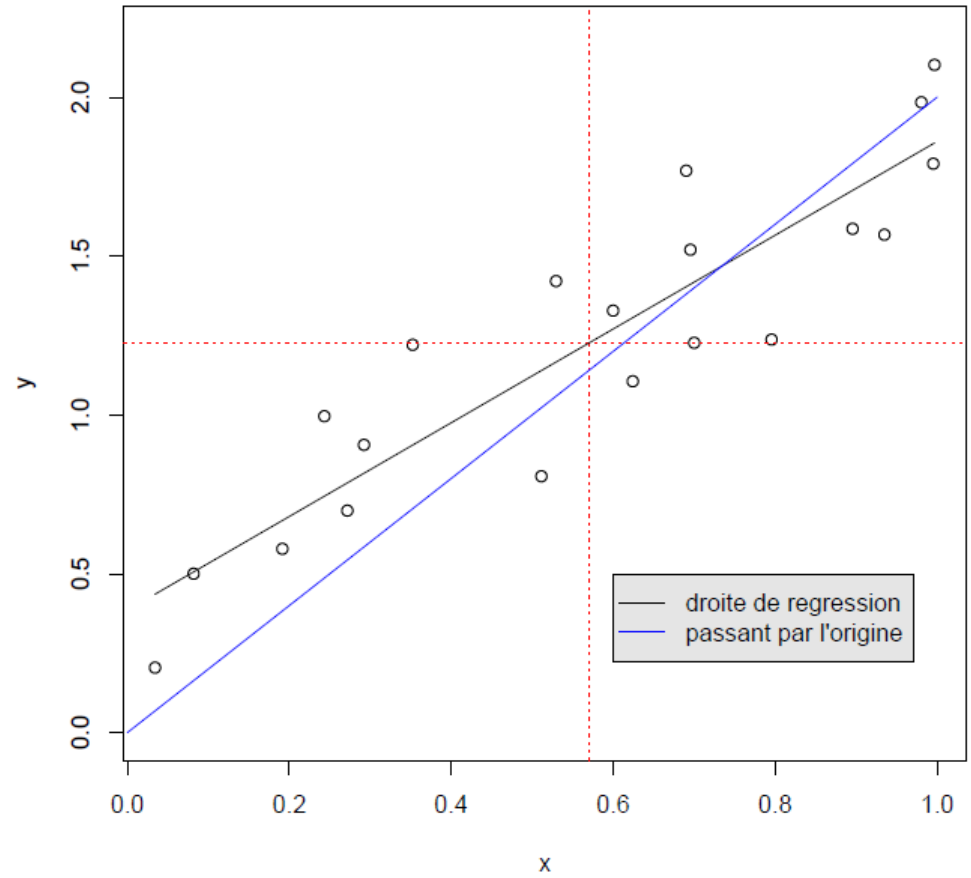
- La régression de **Y par X** n'est pas la même chose que la régression de X par Y



La régression : cas particulier

- Condition technique : (0,0) seul point « certain »
  - Exemple : mesure de spectrophotométrie.

➔Régression : droite d'équation : y = ax (un seul paramètre)



La régression : les points extrêmes

- Extrême sur Y : ordonnée très différente des autres points d'abscisse proche

➔ **Point non consistant**

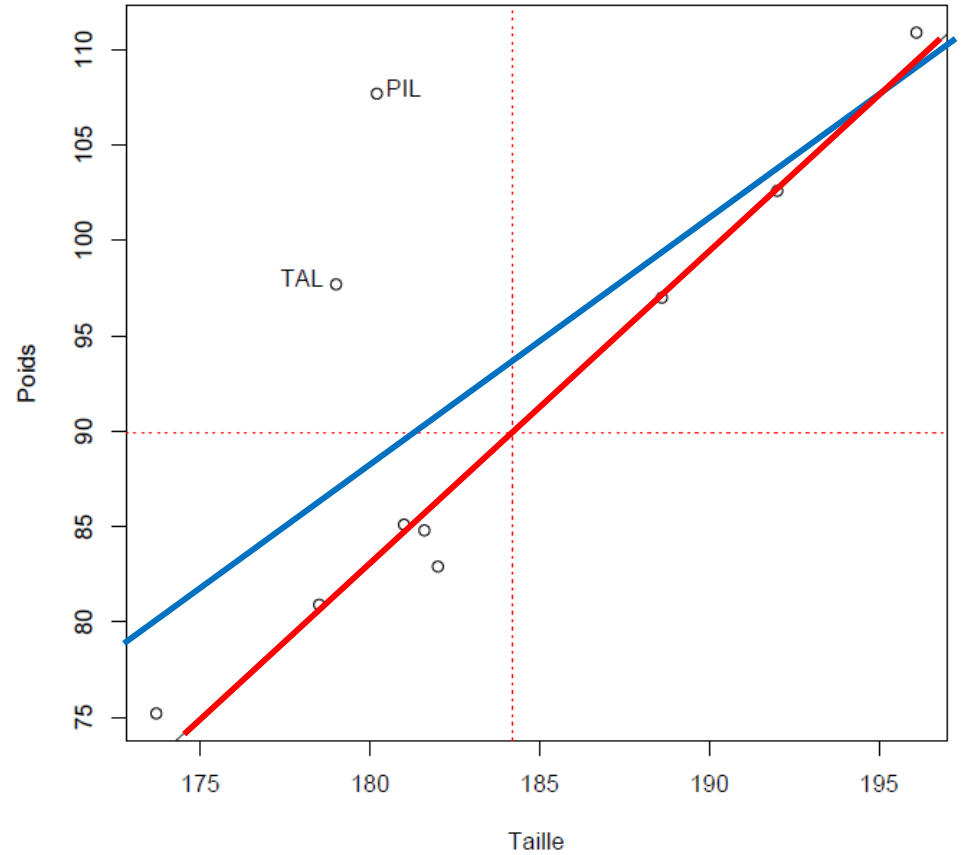
- Extrême sur X : abscisse nettement plus petite ou plus grande que celle des autres points

➔ **Phénomène de levier**

- Un point est **influent** lorsque la régression pratiquée avec ou sans ce point conduit à des résultats très différents.

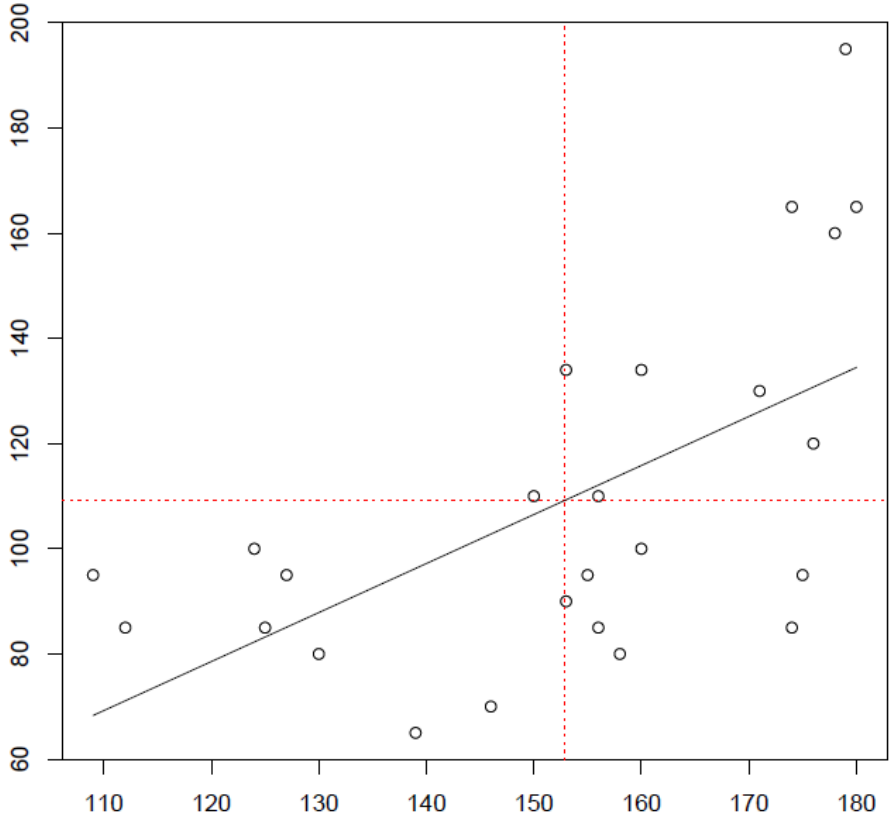
La régression : les points extrêmes

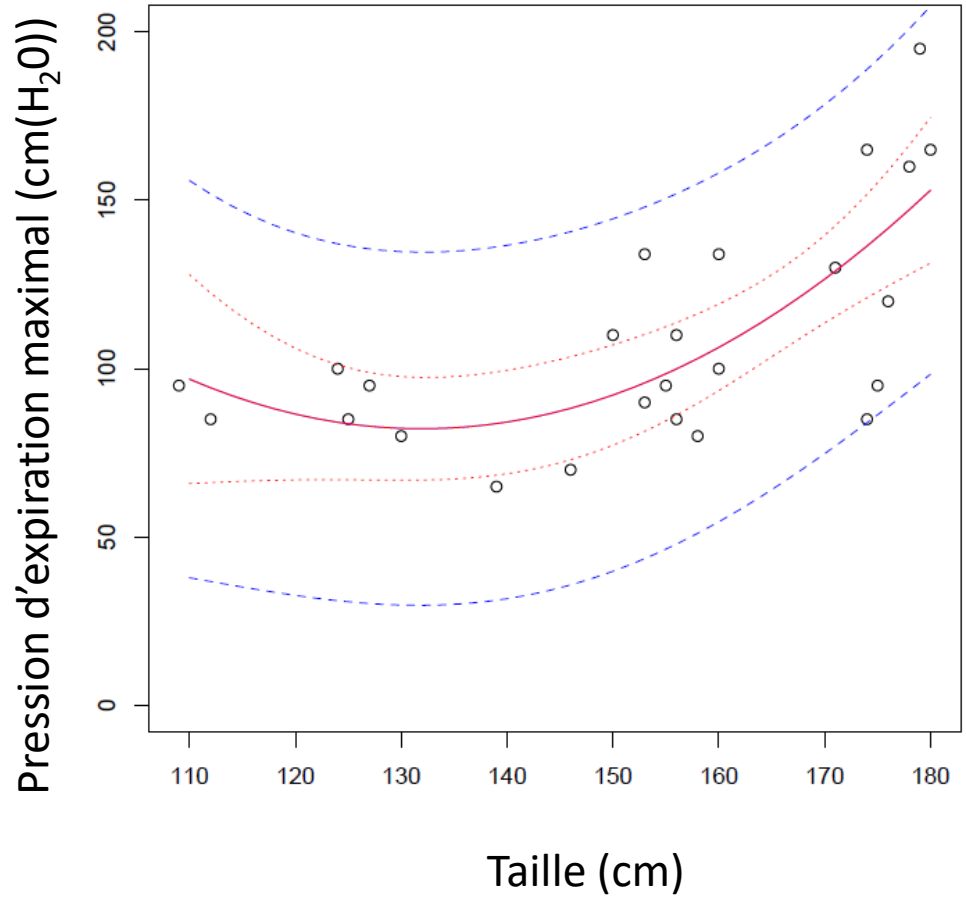
- Tous les rugbymens : $r^2 = 0,51$
- Sans piliers et talonneurs : $r^2 = 0,98$



La régression non linéaire

- P = - 33,28 + 0,93 * taille
- P = 615,36 - 8,08 * taille + 0,06 * taille²





Description bivariée

Deux variables qualitatives

Corrélation entre deux variables qualitatives

- Deux variables qualitatives sont corrélées = un des groupes créé par l'intersections est sur ou sous représenté par rapport aux autres

	Modalité A	Modalité B
Modalité 1	30	12
Modalité 2	15	14

Exemple d'un tableau de contingence

MQDS 2023
Measurement - Quality - Data Science

Visualisation

Carte des points chauds (heat map) ou gauffre

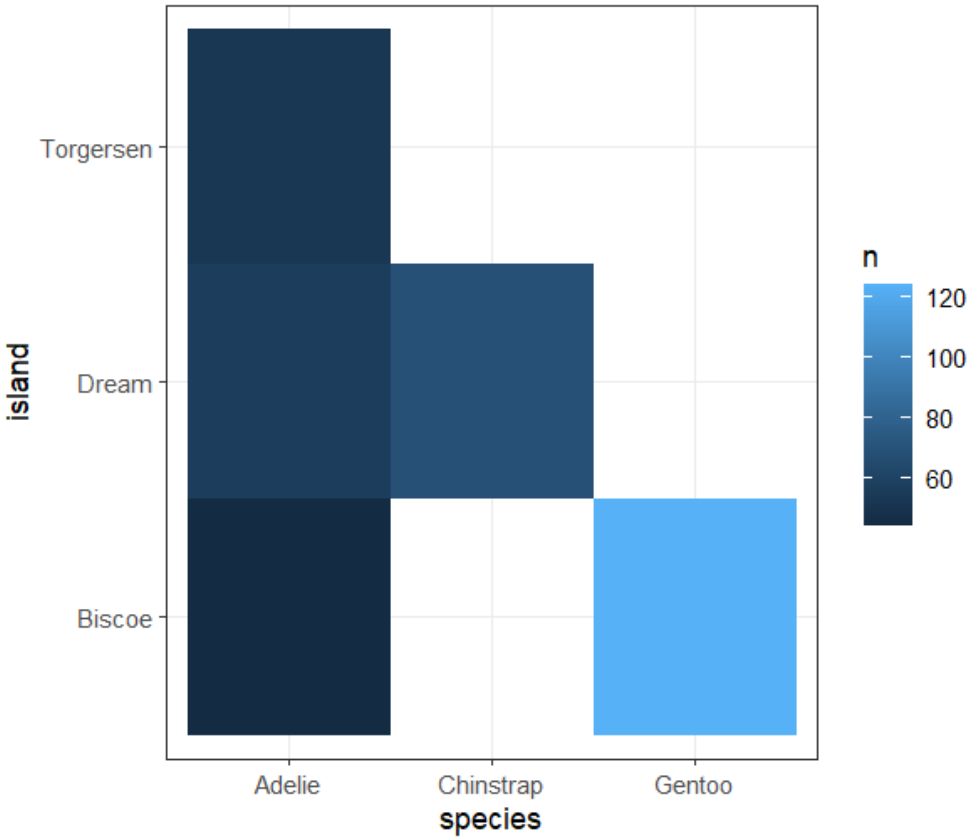
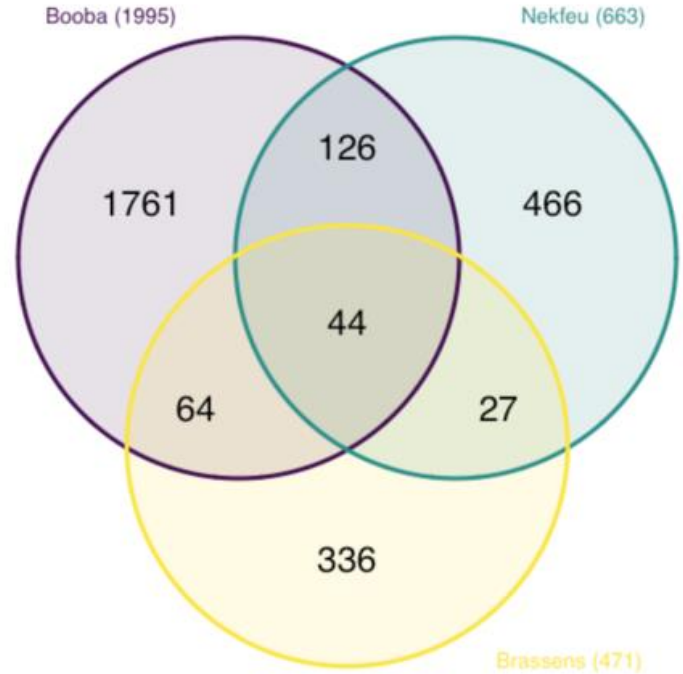


Diagramme de Venn

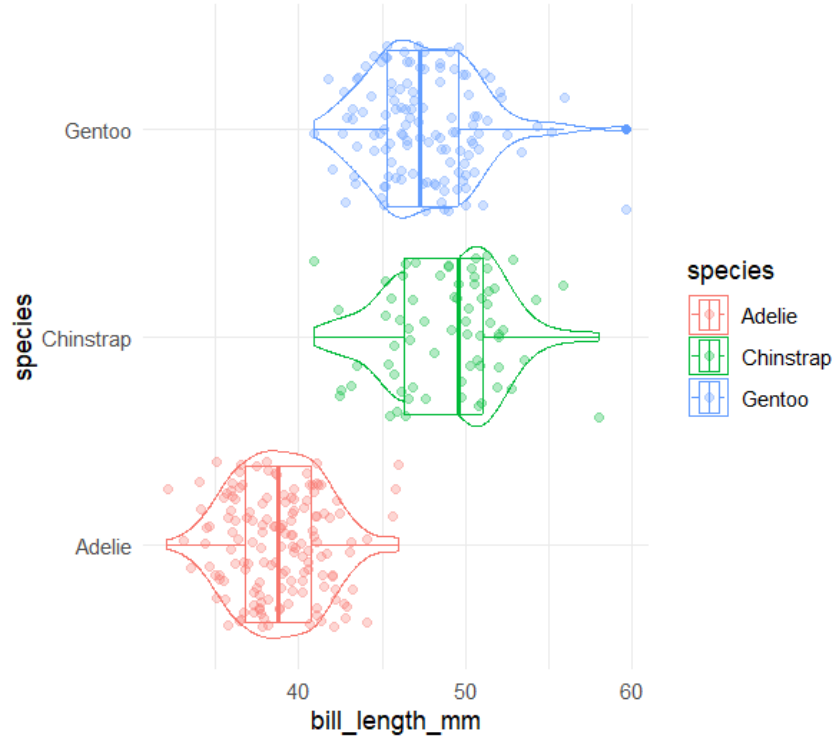


Description bivariée

Une variable qualitative et une quantitative

Différence entre les groupes

L'analyse descriptives bivariées entre une variables qualitative et une quantitative revient à cherche si un des groupes est différents des autres



MQDS 2023
Measurement - Quality - Data Science

TEST PARAMÉTRIQUE :

L'ANOVA
(ANALYSE DE VARIANCES)

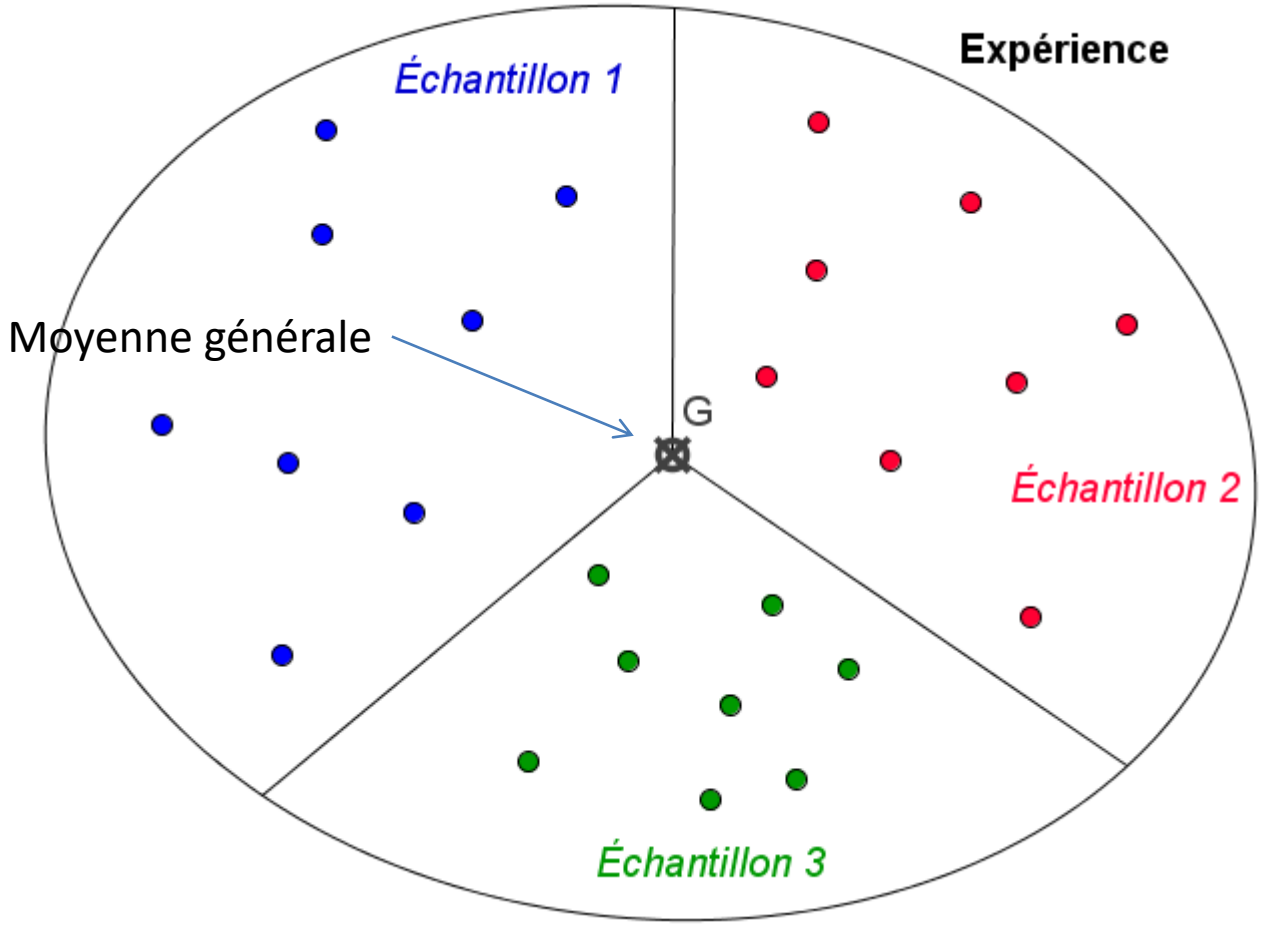
ANOVA : conditions

- Analysis Of Variance
- Test si trois échantillons ou plus sont issus d'une même population (même moyenne)

- Conditions :
  - Échantillonnage aléatoire
  - Indépendance des groupes
  - Normalité des données au sein de chaque groupe (ou approximation par une loi normale si n > 30)
  - Égalité des variances entre les groupes

ANOVA : principe

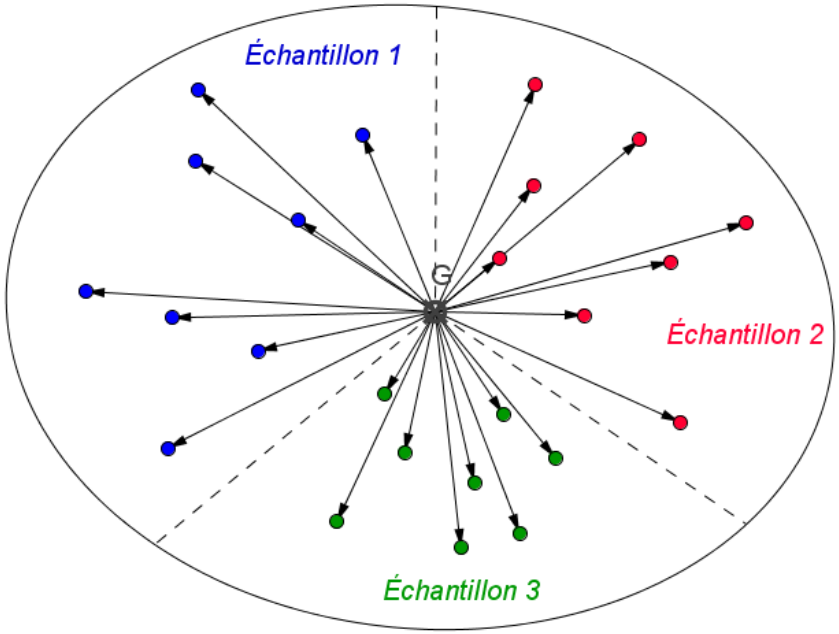
Groupe : variable qualitative



Variable réponse	Groupe
y_1	1
y_2	1
y_3	1
y_4	1
y_5	1
y_6	1
y_7	1
y_8	1
y_9	2
y_{10}	2
y_{11}	2
y_{12}	2
y_{13}	2
y_{14}	2
y_{15}	2
y_{16}	2
y_{17}	3
y_{18}	3
y_{19}	3
y_{20}	3
y_{21}	3
y_{22}	3
y_{23}	3
y_{24}	3

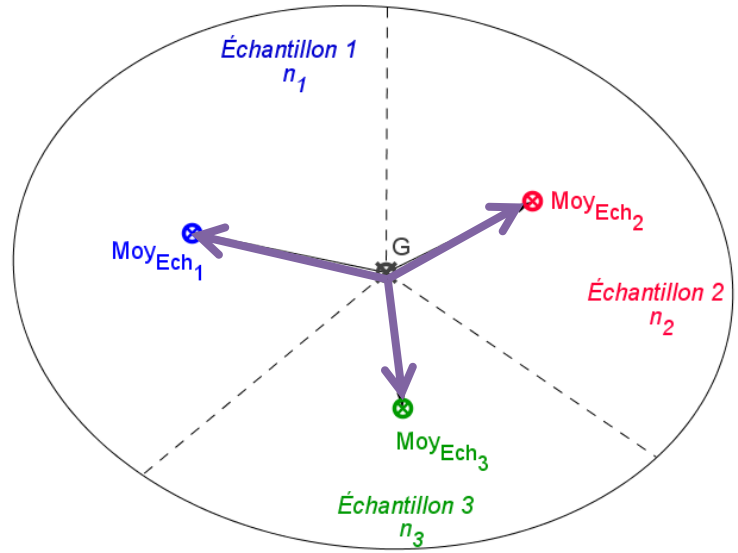
ANOVA : principe

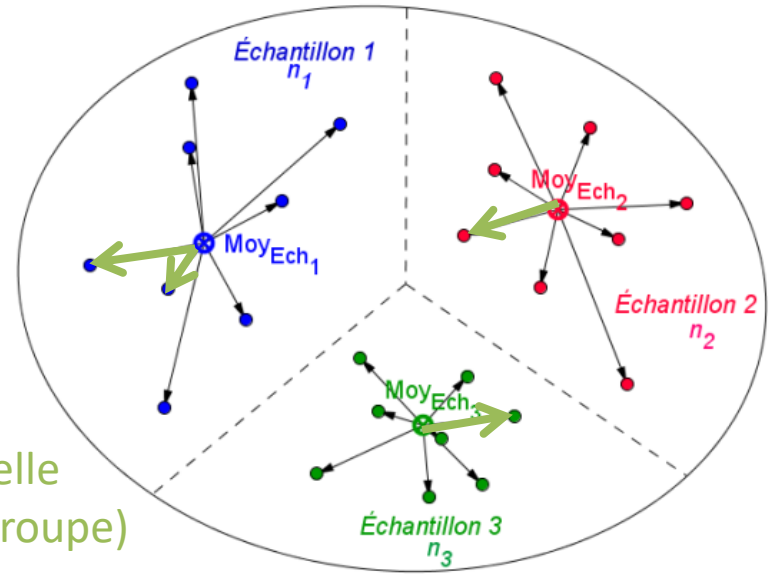
Décomposition de la variabilité



Variabilité totale

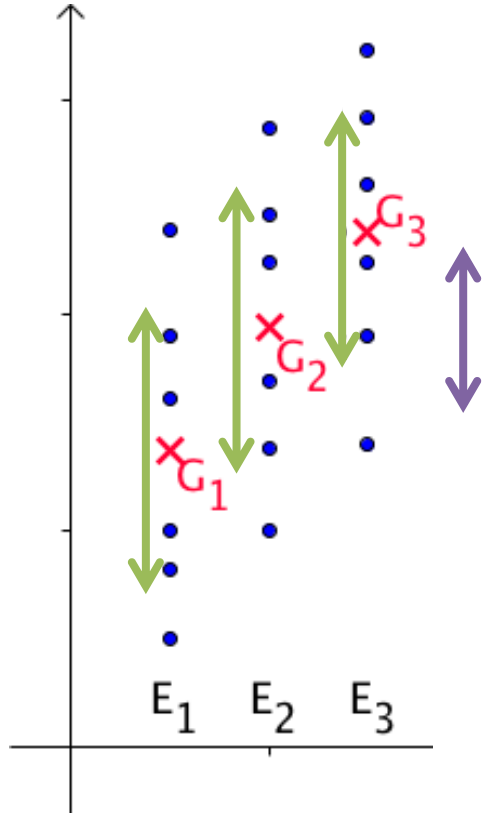
Variabilité factorielle
(entre groupes)



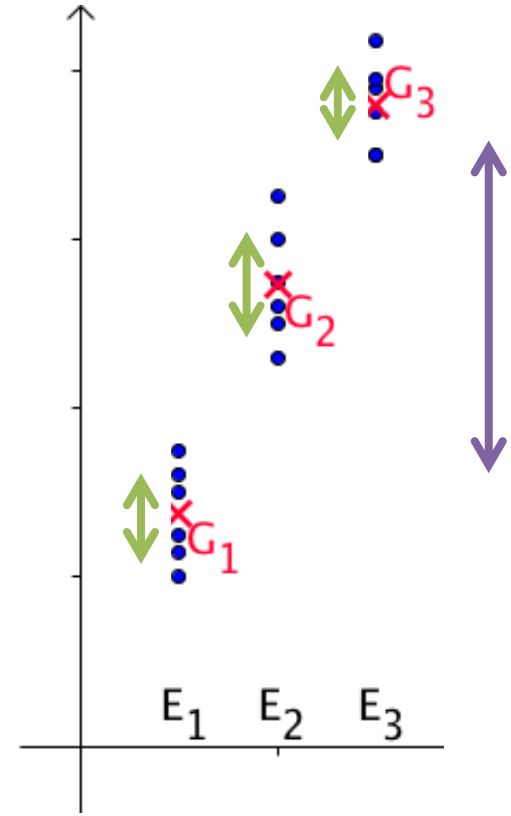


Variabilité résiduelle
(au sein de chaque groupe)

ANOVA : principe



Variabilité factorielle faible
Variabilité résiduelle forte
→ Pas de différence significative



Variabilité factorielle forte
Variabilité résiduelle faible
→ Différence entre les groupes

ANOVA : statistique de test

- H_0 : moyennes similaires entre les groupes
- H_1 : une ou plusieurs moyennes sont différentes

Variation	Somme des Carrés des Ecarts	Degré De Liberté (k : nbr de groupes)	Carré Moyen (variance estimée)
Factorielle	$SCE_F = (k-1)*\sigma_F^2$	k-1	$CM_F = \sigma_F^2$
Résiduelle	$SCE_R = (n-k)*\sigma_R^2$	n-k	$CM_R = \sigma_R^2$
Totale	**$SCE_T = (n-1)*\sigma^2$**	**n-1**	**$CM_T = \sigma^2$**

- $F_{Calculé} = \frac{\sigma_F^2}{\sigma_R^2} = \frac{SCE_F / (k-1)}{SCE_R / (n-k)} = \frac{\text{Variance inter-échantillons}}{\text{Variance intra-échantillons}}$
- F suit une loi de Fisher-Snedecor à (k-1, n-k) DDL
- Si $F_{Calculé} > F_{seuil}$ (k-1, n-k) → rejet de H_0

ANOVA : le test post-hoc

- Si H_0 rejeté : au moins un des groupes est différent mais identité inconnu

➔ Méthode informelle : Création de boites à moustache (boxplot) pour déterminer les différences

➔ Utilisation d'un test post-hoc pour savoir quel(s) groupe(s) sont différent(s) des autres

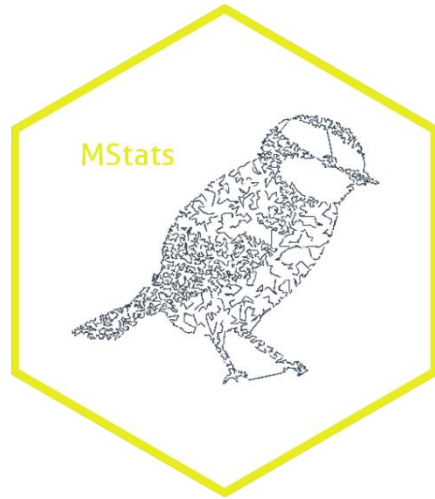
Ex. Test post-hoc de Tuckey (ou test HSD : Honestly Significant Différence)

ANOVA : démarche de test

- Définir les hypothèses H_0 et H_1
- Vérifier que les conditions soient respectées
- Calculer la p-value
  - Si p-value > α → Non rejet de H_0
  - Si p-value ≤ α → Rejet de H_0

  Utilisation d'un test post-hoc pour déterminer les différences entre groupes
- Conclure

Merci pour votre attention !



MQDS 2023
Measurement - Quality - Data Science

Vocabulaire

Définitions (2)

- **Données** : informations, caractéristiques connues, mesurées sur une population.

- **Population** : ensemble cohérent d'éléments homogènes (animaux, plantes, entreprises, clients, boîte de Petri, productions…).

- **Individus statistiques** : unités observées issues de la population, c'est-à-dire entité entière sur laquelle les données sont ou pourraient être récoltées. Les individus statistiques peuvent être des forêts ou des arbres, des troupeaux ou un individu, cela dépend de la question posée.

Définitions (3)

- **Observation** : mesures ou informations récoltées sur les individus statistiques faisant partie de l'échantillon.
- **Échantillon** : ensemble d'individus statistiques sélectionnés faisant partie d'une population.
- **Modalités** : différentes valeurs prises par une variable.

Définitions (4)

- **Plan d'échantillonnage** : organisation des prélèvements à réaliser sur une population. Le plan d'échantillonnage peut être :
  - aléatoire simple : prélèvements d'individus au hasard ;
  - aléatoire systématique : tous les X individus ;
  - aléatoire stratifié : échantillonnage reprenant la structure de la population.

- **Plan d'expérience** : construit en fonction de la question posée et des conditions de prélèvement. Il doit inclure la taille de l'échantillon, la fréquence de toutes les observations, le nombre de réplicas, si la collecte est en milieu naturel, contrôlé, semi-contrôlé ou artificiel… Le plan d'expérience doit être défini en fonction du type de questions posées :
  - Améliorer les connaissances sur un sujet. Par exemple : comment la couverture corallienne évolue-t-elle sur la côte nord-est de l'Australie ?
  - Mieux comprendre un phénomène connu. Par exemple : quels sont les principaux facteurs expliquant la dégradation de la couverture corallienne de la grande barrière de corail ?
  - Prendre une décision éclairée. Par exemple : sur quels facteurs agir pour limiter l'érosion corallienne ?
  - Faire des prédictions. Par exemple : quel serait le taux de couverture corallienne dans 10, 20 ou 50 ans ?

MQDS 2023
Measurement - Quality - Data Science

La collecte de données

- En situation contrôlées (plan d'expérience) ou naturelles
- Données **représentatives** issues d'un plan d'échantillonnage adéquat (aléatoire)
- En **quantité suffisante** : Compromis entre les coûts d'acquisition des données (temps, argent…) et la précision de la réponse