TP OUMOBIO 7: Décrire la relation statistique entre deux variables quantitatives

Mathieu Brevet

2025-01-14

Bienvenue dans ce septième TP sur R. Après avoir vu comment décrire graphiquement les relations entre deux variables statistiques, nous allons plus particulièrement nous attarder sur comment décrire statistiquement la relation entre deux variables quantitatives.

```
setwd("~/ATER PAU 2024/Cours modifiés/OUMOBIOS")


data_lezard = read.table("Suivi_lezard_vivipare.csv", header = T, sep = "\t", dec = ",")
```

Décrire la distribution d'une variable quantitative

Il est souvent très important en statistique d'identifier la distribution théorique (loi de probabilité) pouvant le mieux décrire une variable donnée. Cela est notamment requis pour utiliser des tests statistiques dit **"paramétrique"**, c'est à dire valide uniquement si la variable suit bien un loi de distribution donnée.
Dans notre cas, l'étude de la distribution d'une variable continue peut donner des indications sur la qualité de nos outils de modélisation et sur de potentielles piste de transformation de variable (voir TP suivants). Pour la plupart des mesures continues et finies la loi de distribution usuelle est la loi normale.

La loi de distribution normale

Il est possible dans R de **simuler** des valeurs suivant une **loi de de probabilité** donnée. Ces simulations se réalisent en utilisant les fonctions de la forme **rmaloi** (en remplaçant "maloi" par le nom de la distribution souhaitée: par "norm" pour une loi normale, "binom" pour une loi de Bernoulli, "pois" pour une loi de Poisson, etc.). On peut également obtenir la densité, la fonction de répartition et les quantiles de la distribution à l'aide des fonctions de la forme "dmaloi" (valeur de la fonction de densité à un point donnée, i.e. probabilité de la variable à un point précis pour les lois discrètes), "pmaloi" (valeur de la fonction de répartition à un point donnée, i.e. probabilité que la variable soient inférieure à la valeur donnée), "qmaloi" (quantiles de la loi de probabilité). Ces fonctions peuvent être très utile pour comparer une distribution existante à une distribution théorique ou tout simplement pour simuler une distribution théorique. Nous allons ici prendre un exemple visant à réaliser une simulation de loi normale:

```
# étudions la distribution des masses de lézards:

mean(data_lezard$M_IND)
```

```
## [1] 0.1576786
```

```r
var(data_lezard$M_IND)
```

```
## [1] 0.0003305068
```

```r
sd(data_lezard$M_IND)
```

```
## [1] 0.01817985
```

```r
# sd et var sont des estimateurs non biaisés de l'écart-type et de la variance


# Pour info: démonstration expérimentale du biais de l'estimateur de la variance
var(data_lezard$M_MOTHERS) # variance de la "population"
```

```
## [1] 0.3211427
```

```r
mean(replicate(1000, var(data_lezard$M_MOTHERS[sample(168, 80)])))
```

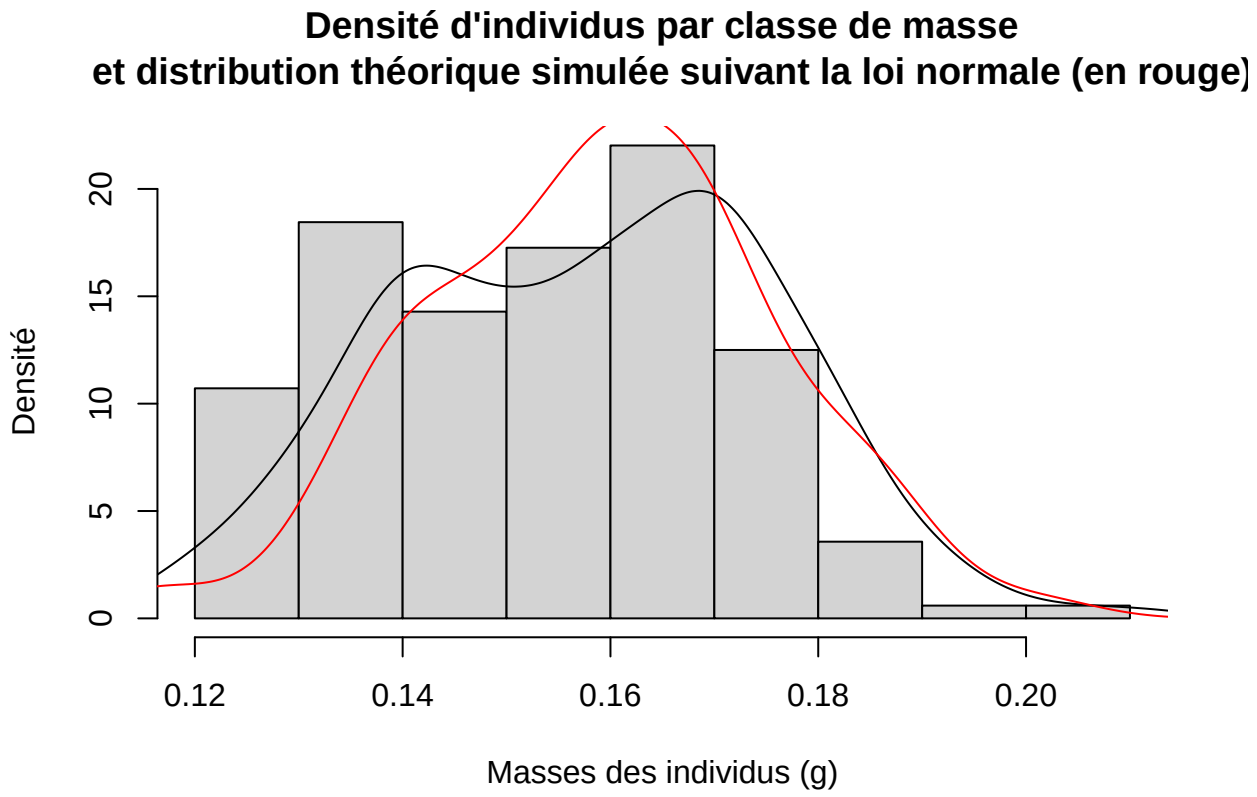
```
## [1] 0.321613
```

```r
# variance moyenne d'un grand nombre d'échantillons de la population,
# avec un estimateurs non biaisé ==> pas de biais
mean(replicate(1000, 79/80 * var(data_lezard$M_MOTHERS[sample(168, 80)])))
```

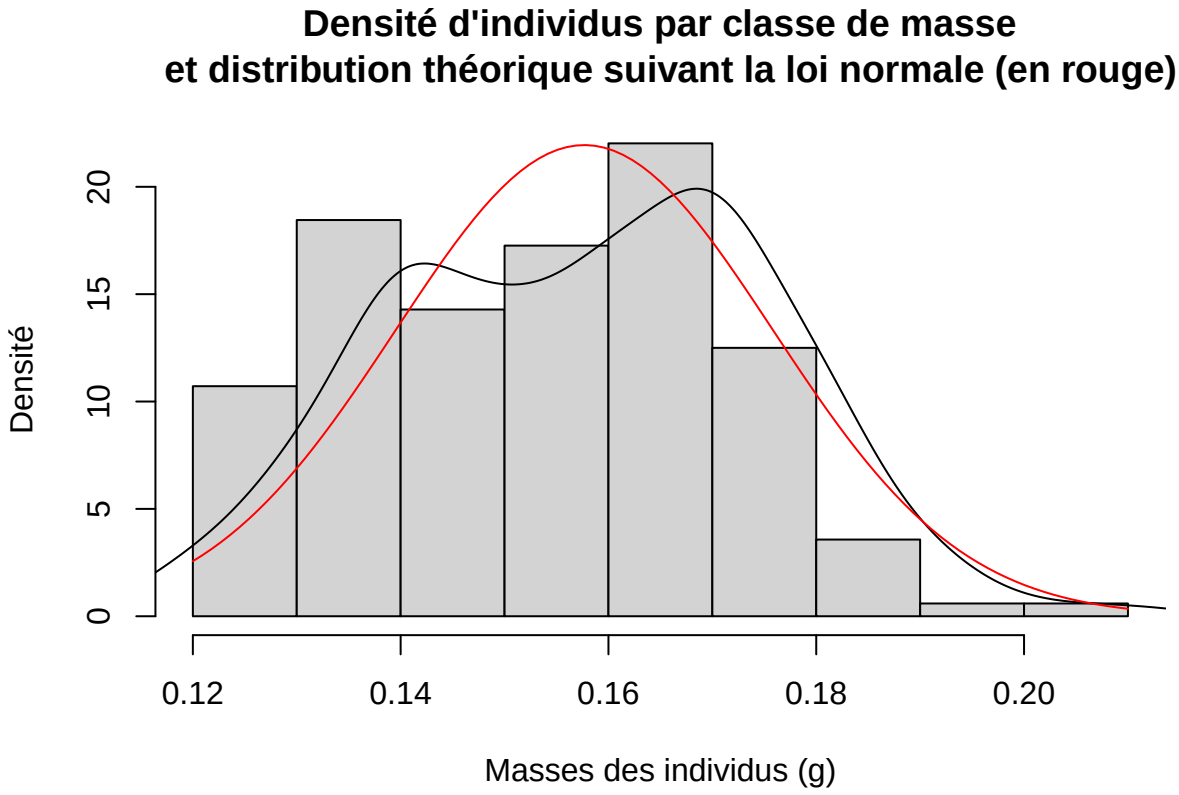
```
## [1] 0.3180879
```

```r
# variance moyenne d'un grand nombre d'échantillons de la population,
# avec un estimateurs biaisé   ==> biais dans l'estimation obtenue




# nous allons simuler un loi de distribution cherchant à se rapprocher de la distribution
# des masses de lézard:

sim_norm = rnorm(length(data_lezard$M_IND), mean(data_lezard$M_IND), sd(data_lezard$M_IND))
# simulation contenant le même nombre d'individus que dans le jeu de données, et dont les
# paramètres de distribution (moyenne, écart-type) sont les mêmes que la variable étudiée


# nous allons graphiquement comparer les deux distributions (théoriques et observés):

hist(
  data_lezard$M_IND,
  prob = T,
  main = "Densité d'individus par classe de masse
  et distribution théorique simulée suivant la loi normale (en rouge)",
  xlab = "Masses des individus (g)",
  ylab = "Densité"
  )
```

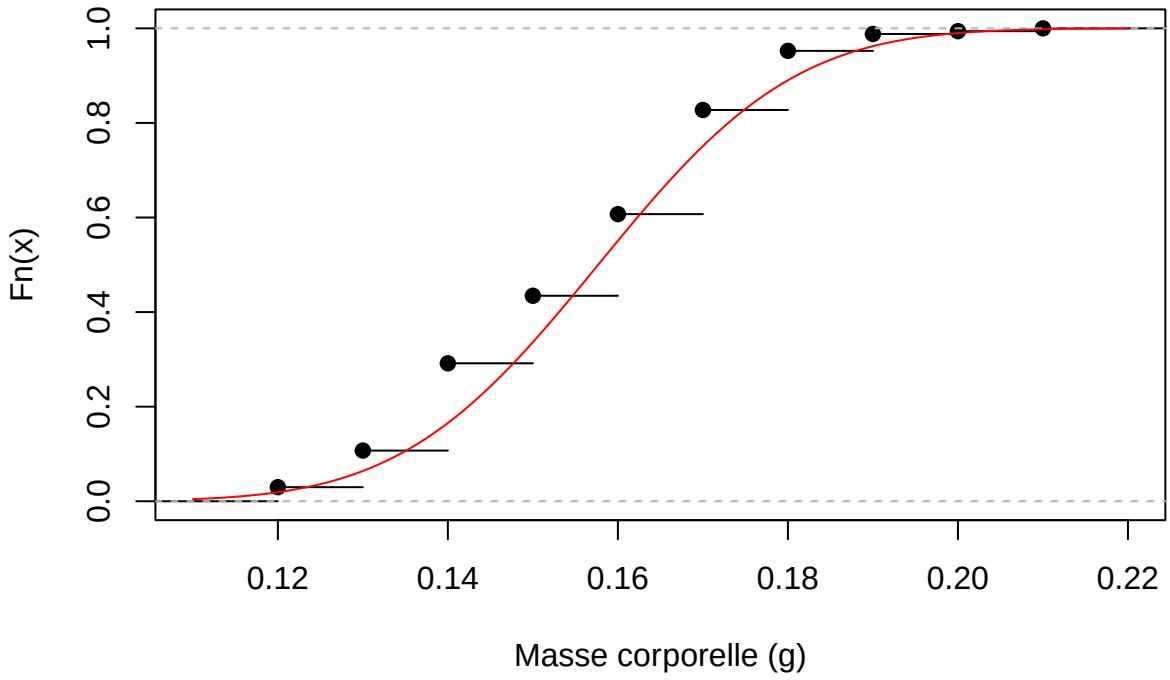
```r
# histogramme, avec une modification des valeurs d'effectifs en ordonnée telle que
# l'aire sous la courbe soit égale à 1 (argument "prob = T")
lines(density(data_lezard$M_IND))
# la fonction "lines()" permet de'ajouter une courbe suivant un tracé donné sur un graphique
lines(density(sim_norm), col = "red")
```




```r
# les deux distributions sont assez proches
# NB: l'argument "col" permet de choisir la couleur du figuré principal de la figure /
# insérer un retour à la ligne dans du texte permet de faire apparaître un retour à la ligne
# (on peut aussi insérer la chaîne de caractère "\n" pour cela)


hist(
  data_lezard$M_IND,
  prob = T,
  main = "Densité d'individus par classe de masse
  et distribution théorique suivant la loi normale (en rouge)",
  xlab = "Masses des individus (g)",
  ylab = "Densité"
  )
lines(density(data_lezard$M_IND))
curve(dnorm(x, mean(data_lezard$M_IND), sd(data_lezard$M_IND)), add=T, col = "red")
```




```
# alternative: pas de simulation, tracé direct de la fonction de densité théorique
# NB: la fonction "curve()" permet de tracer une courbe suivant une fonction donnée
# "x" est l'étendue de la fenêtre graphique actuelle
# "add = T" permet d'ajouter le graphique tracé sur le dernier graphique créé


plot(ecdf(data_lezard$M_IND),
     main="Fonction de répartition observée et théorique (en rouge)",
     xlab = "Masse corporelle (g)")
curve(pnorm(x, mean(data_lezard$M_IND), sd(data_lezard$M_IND)),
      add=T,
      col="red")
```



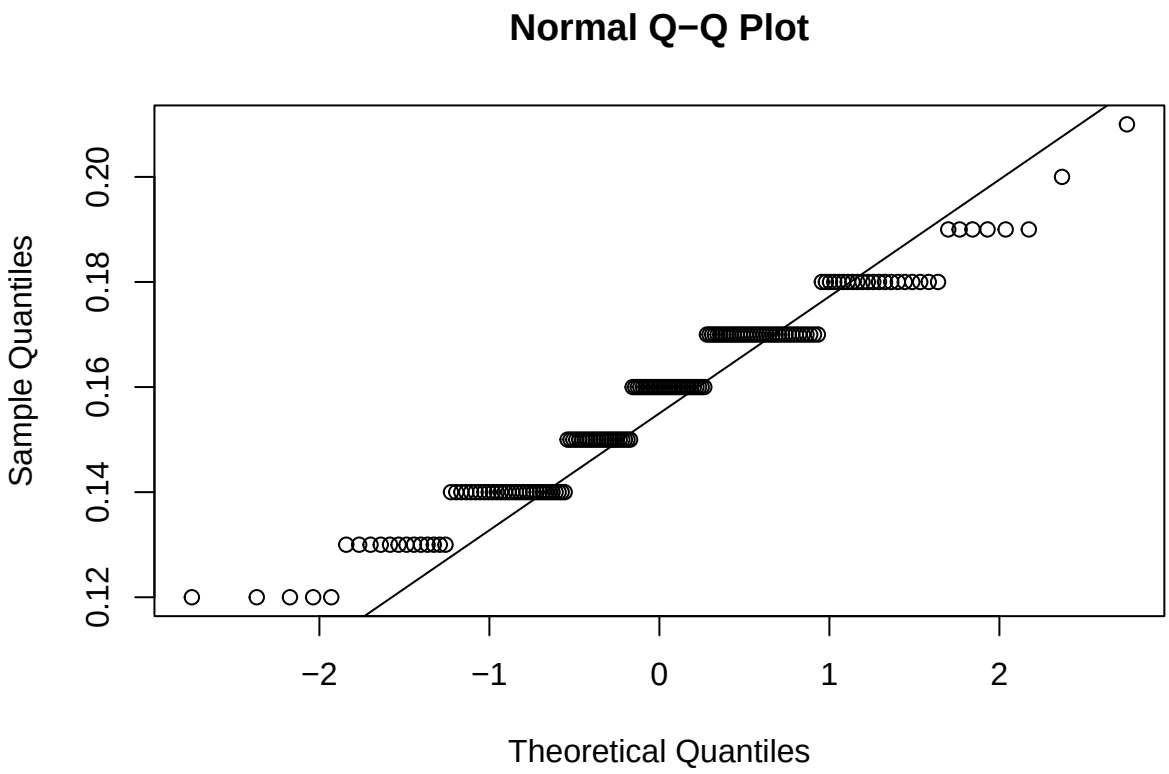

```
# la fonction "ecdf()" permet de tracer la fonction de répartition empirique
# à partir des données fournies




# la comparaison peut également se faire sur les quantiles:


qnorm(c(0.25, 0.5, 0.75), mean(data_lezard$M_IND), sd(data_lezard$M_IND))
```

```
## [1] 0.1454165 0.1576786 0.1699407
```

```
summary(data_lezard$M_IND)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.1400  0.1600  0.1577  0.1700  0.2100
```

```
qqnorm(data_lezard$M_IND)
# graphe de comparaison des quantiles théoriques et observés
qqline(data_lezard$M_IND)
```




```
# droite théorique que devraient suivre les valeurs en cas d'adéquation avec la loi normale


# l'adéquation à une loi normale est plutôt bonne même si on observe des déviations
# sur les queues de distribution
```

EXERCICE

- Étudiez graphiquement l'adéquation de la distribution des masses corporelles des mères à une loi normale. Comment pouvez-vous estimer les paramètres de la loi normale approchant cette distribution ?
- Donnez l'intervalle de confiance à 95% des masses corporelles des mères. Faites de même pour les masses des mères de strictement moins de 60mm et de strictement plus de 65mm, que pouvez vous conclure ?

```
hist(
  data_lezard$M_MOTHERS,
  prob = T,
  main = "Densité d'individus par classe de masse
  et distribution théorique suivant la loi normale (en rouge)",
  xlab = "Masses des mères (g)",
  ylab = "Densité"
  )
```

```r
lines(density(data_lezard$M_MOTHERS))
curve(dnorm(x, mean(data_lezard$M_MOTHERS), sd(data_lezard$M_MOTHERS)), add=T, col = "red")
```

```r
qqnorm(data_lezard$M_MOTHERS)
qqline(data_lezard$M_MOTHERS)
```

```r
# l'adéquation à une loi normale est plutôt bonne même si on observe des déviations
# sur les queues de distribution



# simulation d'une loi normale approchant la distribution étudiée:

sim_norm = rnorm(length(data_lezard$M_MOTHERS),
                 mean(data_lezard$M_MOTHERS),
                 sd(data_lezard$M_MOTHERS))




# intervalle de confiance à 95% de la masse des mères:

IC_M_MOTHERS = c(mean(data_lezard$M_MOTHERS) - 1.96 * sd(data_lezard$M_MOTHERS),
                 mean(data_lezard$M_MOTHERS) + 1.96 * sd(data_lezard$M_MOTHERS))




# intervalle de confiance à 95% de la masse des juvéniles mesurant moins ou plus de 20mm:


# il faut tout d'abord vérifier que les distributions étudiées peuvent bien être aprroximé
# par une loi normale:

qqnorm(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS)
qqline(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS)
```

```r
qqnorm(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS)
qqline(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS)
```

```r
# l'adéquation à une loi normale est discutable dans les deux cas, mais on considérera
# qu'elle est suffisante pour une approximation


IC_M_MOTHERS_inf_60 =
  c(mean(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS) -
      1.96 * sd(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS),
    mean(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS) +
      1.96 * sd(data_lezard[data_lezard$SVL_MOTHERS < 60,]$M_MOTHERS))
```

```r
IC_M_MOTHERS_sup_65 =
  c(mean(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS) -
      1.96 * sd(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS),
    mean(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS) +
      1.96 * sd(data_lezard[data_lezard$SVL_MOTHERS > 65,]$M_MOTHERS))

# Les deux intervalles de confiance ne se recoupent pas, cela signifie qu'au moins 95%
 # des individus de chaque groupe ne peuvent pas avoir le même poids. Il y a donc très
# peu de probabilité d'avoir des individus de chaque groupe ayant le même poids puisqu'une
# très grande majorité  des poids de chaque groupe sont distribués dans des intervalles
# disjoints.
```

NOTE IMPORTANTE

Dans le cas où différentes séries de valeurs continues suivent une loi normale et sont comparables (mesures de la même métrique, dans la même unité), on peut utiliser leurs intervalles de confiance à 95% pour déterminer si la métrique étudiée diffère entre les différents groupes étudiés. En effet, si leurs intervalles de confiance sont dijoints (pas de chevauchement) il existe une faible probabilité d'avoir des individus provenant des différents groupes partageant la même valeur pour la métrique étudiée.

Décrire la relation statistique entre deux variables quantitatives

Les relations entre deux variables quantitatives peuvent être étudiés à l'aide de plusieurs métriques statistique. La première est la **covariance**, elle permet de savoir comment les deux variables co-varient et dans quel sens, cette mesure est sensible à l'unité des variables (elle est dite dimensionnelle). La seconde est la **corrélation**, il s'agit d'une mesure normalisé de la covariance (qui n'est donc pas sensible aux unités: adimensionnelle) qui permet de renseigner la prépondérance de la relation entre variable sur les variations internes des variables.
Dans le cas d'une relation linéaire, on utilisera la **corrélation de Pearson** (relation affine entre variable), dans le cas d'une relation non linéaire mais monotone (croissante ou décroissante) on pourra utiliser la **corrélation de Spearman**. La corrélation de Spearman peut aussi être utilisée pour étudier le lien entre des variables quantitative discrètes et des variables qualitatives ordinales (rang, classement...).

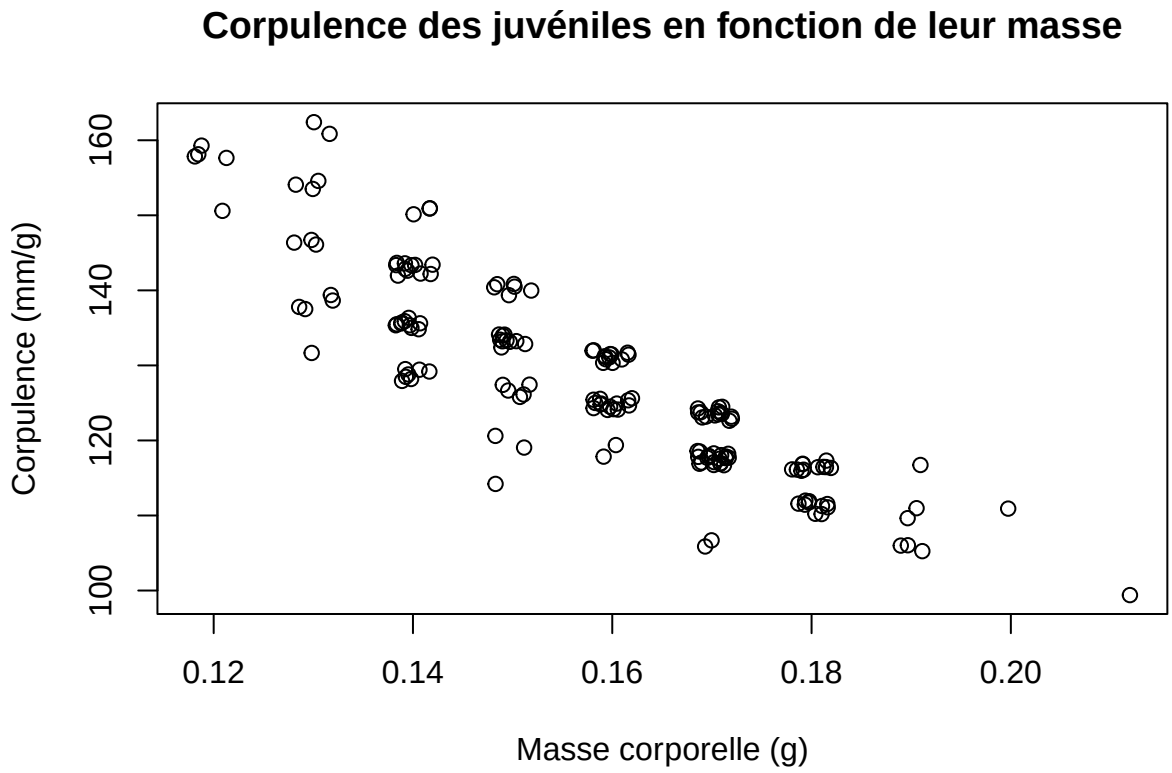
NOTE IMPORTANTE

Les mesures statistique réalisées ici informent simplement sur la nature de la relation entre deux variable (croissante/décroissante, existence ou non d'une association et force de l'association entre deux variables) mais ne disent rien sur les paramètres de cette relation (voir TP suivant), ni sur l'existence d'une relation de causalité entre variables. Une corrélation importante peut tout à fait être due au hasard ou être entrainé par des relations causales avec des variables tiers. De nombreux exemple de corrélations "fallacieuse" peut être trouvés ici par exemple: https://www.tylervigen.com/spurious-correlations

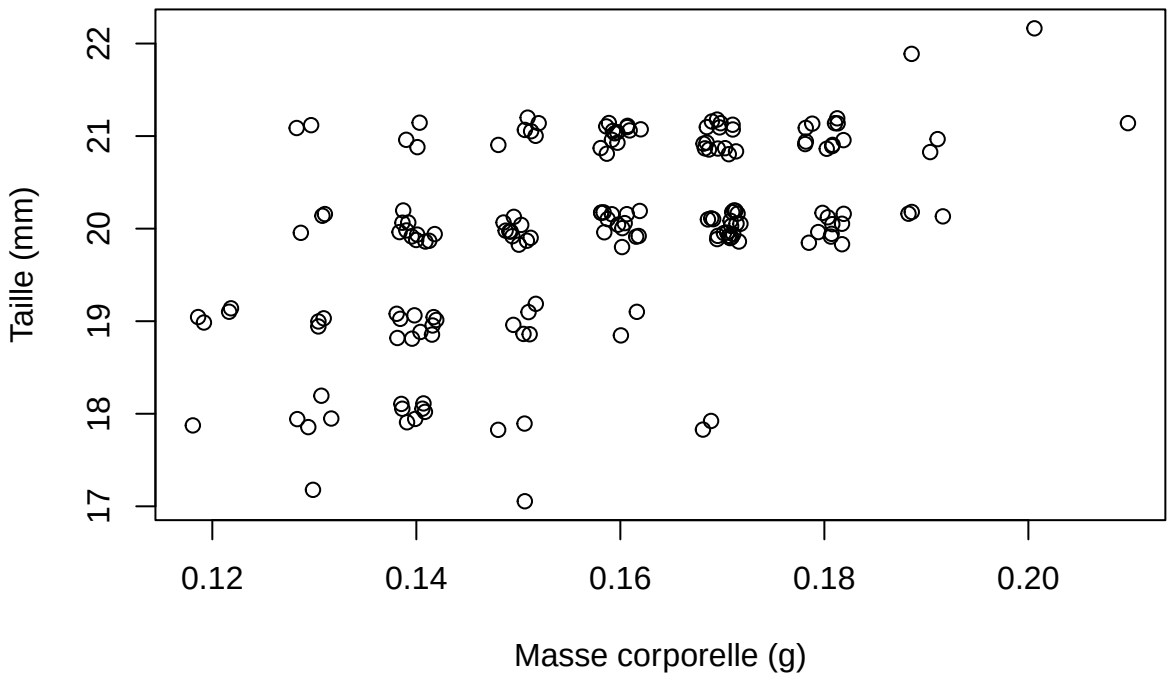
Nous allons prendre ici deux exemples de relation entre variable quantitative, linéaire et monotone:

```r
plot(
    jitter(data_lezard$SVL_IND/data_lezard$M_IND, 10) ~
        jitter(data_lezard$M_IND, 1),
```

```
    main = "Corpulence des juvéniles en fonction de leur masse", xlab = "Masse corporelle (g)",
    ylab = "Corpulence (mm/g)"
)
```




```
# relation linéaire entre corpulence et masse des juvéniles



plot(
    jitter(data_lezard$SVL_IND, 1) ~
        jitter(data_lezard$M_IND, 1),
    main = "Taille des juvéniles de lézard en fonction de leur masse", xlab = "Masse corporelle (g)",
    ylab = "Taille (mm)"
)
```




```
# relation monotone (croissante) mais pas forcément linéaire entre taille et
# poids des individus




# Calcul covariance et corrélation:


cov(data_lezard$SVL_IND/data_lezard$M_IND, data_lezard$M_IND)  # covariance
```

```
## [1] -0.2082934
```

```
cor(data_lezard$SVL_IND/data_lezard$M_IND, data_lezard$M_IND)  # par défaut corrélation de
```

```
## [1] -0.8949622
```

```
# Pearson (adaptée au relation linéaire) Forte relation (valeur du coefficient
# de corrélation proche de -1) décroissante (signe négatif de la covariance et
# de la correlation) entre les deux variables

cov(data_lezard$SVL_IND, data_lezard$M_IND)
```

```
## [1] 0.009870616
```

```
cor(data_lezard$SVL_IND, data_lezard$M_IND, method = "spearman")  # corrélation de Spearman
```

```
## [1] 0.5259746
```

```
# (adaptée au relation non-linéaire, monotone) Relation croissante (signe
# positif de la covariance et de la correlation) d'assez faible niveau (valeur
# du coefficient de corrélation proche de 0.5)




# Remarque importante: la corrélation de Spearman se calcule à partir des rangs
# des valeurs pour chaque variable (voir ci-dessous pur l'obtention des rangs)

rank(data_lezard$SVL_IND)
```

```
##   [1]  78.0  30.5 140.0 140.0  30.5  30.5 140.0  78.0  78.0  78.0  30.5  78.0
##  [13] 140.0  78.0  30.5 140.0  78.0  78.0  30.5 140.0 140.0  30.5 140.0  78.0
##  [25]  78.0 140.0 140.0  30.5  78.0  78.0  78.0 140.0  10.5  10.5  78.0 140.0
##  [37]  78.0  10.5  78.0  78.0   1.5  78.0 140.0  30.5  78.0  78.0  30.5  10.5
##  [49]  78.0  30.5  30.5 140.0 140.0  30.5  78.0 140.0  78.0  78.0  78.0  78.0
##  [61] 140.0  78.0  78.0  30.5  78.0 140.0  78.0  78.0  10.5 140.0  10.5  78.0
##  [73] 140.0  78.0  10.5  30.5 140.0  10.5  78.0 140.0  78.0  10.5  10.5  78.0
##  [85]  78.0  78.0 167.5  10.5 140.0  10.5  30.5 140.0 140.0  30.5  30.5  78.0
##  [97]  78.0 140.0 140.0 140.0  78.0 140.0  10.5  78.0  30.5  78.0 140.0  78.0
## [109] 140.0 140.0  30.5  78.0  10.5 140.0  78.0 140.0 140.0 140.0  78.0  78.0
## [121]  78.0 140.0  78.0  78.0 140.0  78.0 167.5  78.0  78.0 140.0  78.0 140.0
## [133]  78.0 140.0  30.5 140.0  78.0  78.0 140.0  10.5  78.0 140.0 140.0  30.5
## [145]  78.0  78.0  78.0  78.0   1.5  78.0  78.0  78.0  10.5 140.0 140.0  78.0
## [157] 140.0  78.0 140.0  78.0 140.0  30.5  78.0  30.5 140.0 140.0  78.0 140.0
```

```
rank(data_lezard$M_IND)
```

```
##   [1]  88.0  34.0 121.0 150.0  61.5  34.0 150.0 150.0 121.0  88.0  34.0 121.0
##  [13] 150.0 150.0  34.0  88.0  61.5  34.0  34.0 150.0  61.5  34.0 121.0 150.0
##  [25]  88.0 150.0  34.0  88.0  34.0 150.0 121.0  88.0  34.0  34.0 121.0  88.0
##  [37] 121.0  12.0 163.5  61.5  12.0 121.0 150.0  34.0 121.0  61.5  61.5  12.0
##  [49]  12.0  12.0  12.0  88.0  88.0  61.5  88.0 121.0 121.0 121.0  61.5 121.0
##  [61] 121.0 121.0  88.0  34.0  61.5 121.0  34.0  61.5  34.0 121.0  61.5  61.5
##  [73] 121.0  88.0  12.0   3.0 121.0  34.0  88.0  61.5  88.0  34.0  61.5 121.0
##  [85] 121.0  88.0 167.0  34.0  88.0  34.0  88.0  88.0 150.0  34.0  61.5  34.0
##  [97] 150.0 121.0  88.0 121.0  61.5  61.5   3.0 121.0   3.0 150.0  61.5 121.0
## [109] 121.0 121.0  34.0  34.0 121.0  88.0 121.0 121.0 121.0 150.0  12.0 121.0
## [121] 121.0 150.0  88.0  61.5  88.0  88.0 163.5  88.0 150.0  88.0  88.0 150.0
## [133] 121.0 168.0  61.5  12.0 121.0  34.0  88.0  12.0 150.0  34.0  61.5  12.0
## [145]  34.0  88.0  34.0 163.5  61.5 150.0 150.0 163.5 121.0 163.5  88.0  12.0
## [157] 163.5  61.5  61.5  34.0 121.0   3.0  34.0   3.0  12.0  34.0  34.0 150.0
```

```r
# dans le cas où il y a des valeurs ex-eaquo comme ici on attribue la valeur
# moyenne des rangs occupés par les valeurs ex-aequo
```

POUR ALLER PLUS LOIN

Il existe d'autres méthodes permettant de mesurer le degré d'association entre deux variables quantitatives. Le coefficient de corrélation de Kendall par exemple permet de mesurer l'association ordinale de deux variables (corrélation de rang) en se basant sur la concordance/discordance des paires de points observés.

EXERCICE

- Étudiez les relations (graphiquement et statistiquement) entre la corpulence (masse divisée par taille) et le poids des mères, et entre le poids et la taille des mères.

- Convertissez la taille des mères en centimètres, calculez de nouveau la covariance et la corrélation entre le poids et la taille des mères, qu'observez-vous ? Pourquoi ?
  Calculez de nouveau ces métriques en centrant-réduisant (soustraction des valeurs par leur moyenne, puis division par leur écart-type) les variables au préalable: quel est l'utilité de cette manipulation ?
- Sachant que la covariance (estimateur non biaisé) est définie par:

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

  Que le coefficient de corrélation de Pearson est défini par:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  Et que le coefficient de corrélation de Spearman est usuellement défini par:

$$\mathrm{Rho}(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

  Reproduisez par vous même les fonctions permettant de calculer ces trois métriques.

- Calculez la corrélation entre la taille et le poids des mères à l'aide votre fonction et de la fonction disponible sur R. Qu'observez-vous ? Faites de même, mais en écartant les valeurs dupliquées de corpulence et de poids: que pouvez vous en conclure ?
- Modifiez votre fonction de covariance pour qu'elle renvoie un message d'erreur si les deux vecteurs en entrée ne font pas la même taille.

```r
plot(
  jitter(data_lezard$SVL_MOTHERS / data_lezard$M_MOTHERS, 10) ~ jitter(data_lezard$M_MOTHERS, 1),
     main = "Corpulence des juvéniles en fonction de leur masse",
     xlab = "Masse corporelle (g)",
     ylab = "Corpulence (mm/g)")
```

```r
# relation linéaire entre corpulence et masse des mères


plot(
  jitter(data_lezard$SVL_MOTHERS, 1) ~ jitter(data_lezard$M_MOTHERS, 1),
     main = "Taille des juvéniles de lézard en fonction de leur masse",
     xlab = "Masse corporelle (g)",
     ylab = "Taille (mm)")
```

```r
# relation monotone (croissante) mais pas forcément linéaire entre taille et poids des mères




cov(data_lezard$SVL_MOTHERS / data_lezard$M_MOTHERS, data_lezard$M_MOTHERS)
cor(data_lezard$SVL_MOTHERS / data_lezard$M_MOTHERS, data_lezard$M_MOTHERS) # corrélation de
# Pearson (adaptée au relation linéaire)
# Forte relation décroissante entre les deux variables

cov(data_lezard$SVL_MOTHERS, data_lezard$M_MOTHERS)
cor(data_lezard$SVL_MOTHERS, data_lezard$M_MOTHERS, method = "spearman") # corrélation de
# Spearman (adaptée au relation non-linéaire, monotone)
# Forte relation croissante




# conversion de la taille des mères en centimètres:

data_lezard$SVL_MOTHERS_cm = data_lezard$SVL_MOTHERS / 10

cov(data_lezard$SVL_MOTHERS_cm, data_lezard$M_MOTHERS)
cor(data_lezard$SVL_MOTHERS_cm, data_lezard$M_MOTHERS, method = "spearman")
# la covariance a dimuné d'un facteur 10 tandis que le coefficient de corrélation a gardé la même valeu
# seule la covariance est sensible à l'unité des variables (métrique dimensionnelle)




# Résultats après avoir centré-réduit les variables:

cov(
  (data_lezard$SVL_MOTHERS - mean(data_lezard$SVL_MOTHERS)) /  sd(data_lezard$SVL_MOTHERS),
    (data_lezard$M_MOTHERS - mean(data_lezard$M_MOTHERS)) / sd(data_lezard$M_MOTHERS))
cor(
  (data_lezard$SVL_MOTHERS - mean(data_lezard$SVL_MOTHERS)) /  sd(data_lezard$SVL_MOTHERS),
    (data_lezard$M_MOTHERS - mean(data_lezard$M_MOTHERS)) / sd(data_lezard$M_MOTHERS),
  method = "spearman")

cov(
  (data_lezard$SVL_MOTHERS_cm - mean(data_lezard$SVL_MOTHERS_cm)) /  sd(data_lezard$SVL_MOTHERS_cm),
```

```r
    (data_lezard$M_MOTHERS - mean(data_lezard$M_MOTHERS)) / sd(data_lezard$M_MOTHERS))
cor(
  (data_lezard$SVL_MOTHERS_cm - mean(data_lezard$SVL_MOTHERS_cm)) /  sd(data_lezard$SVL_MOTHERS_cm),
    (data_lezard$M_MOTHERS - mean(data_lezard$M_MOTHERS)) / sd(data_lezard$M_MOTHERS),
  method = "spearman")

# après avoir centré-réduit les variables on observe plus de différences sur les valeurs
# de covariance lors de changements d'unité, on est passé sur des valeurs adimensionnelle

# solution alternative avec la fonction "scale" permettant de centrer-réduire:
cov(scale(data_lezard$SVL_MOTHERS), scale(data_lezard$M_MOTHERS))
cor(scale(data_lezard$SVL_MOTHERS), scale(data_lezard$M_MOTHERS), method = "spearman")

cov(scale(data_lezard$SVL_MOTHERS_cm), scale(data_lezard$M_MOTHERS))
cor(scale(data_lezard$SVL_MOTHERS_cm), scale(data_lezard$M_MOTHERS), method = "spearman")
```

```r
# Création de fonctions mesurant la covariance/correlation:


myCov = function(x, y) {
  sum((x-mean(x)) * (y-mean(y))) / (length(x) - 1)
}


myCorr = function(x, y) {
  cov(x, y) / sqrt(var(x) * var(y))
}


myRho = function(x, y) {
  1 - 6 * sum((rank(x) - rank(y))^2) / (length(x) * (length(x)^2 - 1))
}
```

```r
cor(data_lezard$SVL_MOTHERS, data_lezard$M_MOTHERS, method = "spearman")
myRho(data_lezard$SVL_MOTHERS, data_lezard$M_MOTHERS)

cor(
  data_lezard[!duplicated(data_lezard$SVL_MOTHERS) & !duplicated(data_lezard$M_MOTHERS),]$SVL_MOTHERS,
    data_lezard[!duplicated(data_lezard$SVL_MOTHERS) & !duplicated(data_lezard$M_MOTHERS),]$M_MOTHERS,
  method = "spearman")
myRho(
  data_lezard[!duplicated(data_lezard$SVL_MOTHERS) & !duplicated(data_lezard$M_MOTHERS),]$SVL_MOTHERS,
      data_lezard[!duplicated(data_lezard$SVL_MOTHERS) & !duplicated(data_lezard$M_MOTHERS),]$M_MOTHERS
  )
```

```r
# notre fonction donne un résultat identique à celle de R uniquement lorsqu'il n'y pas de valeurs ex-ae



# Amélioration fonction covariance:


myCov = function(x, y) {
  if (length(x) != length(y)) {
    stop("Les deux vecteurs n'ont pas la même taille.")
  }
  else {
     sum((x-mean(x)) * (y-mean(y))) / (length(x) - 1)
  }
}
```

NOTE IMPORTANTE

En général on privilégiera l'utilisation de la corrélation sur la covariance puisque cette mesure est adimensionelle, et est plus adaptée pour estimer le niveau de la relation entre variables (fortement liées ou non). Lorsqu'on utilise des mesures de covariance, il peut être utile de centrer-réduire les variables préalablement aux analyses pour s'affranchir des problèmes de dimensions

Bilan

La distribution d'une variable continue, prenant des valeurs finies, peut très souvent être approchée par une **loi de distribution normale** (aussi appelée gaussienne) ayant pour paramètre la moyenne et l'écart-type de la variable. On peut simuler dans R une loi normale de paramètres choisis, à l'aide de la fonction **rnorm**. L'adéquation des données à cette loi de distribution théorique peut être vérifiée graphiquement, à l'aide des fonctions suivantes:

- **hist()** pour tracer l'histogramme des données observées / **lines(density())** pour tracer la courbe de densité des données ou de la loi normale simulée (dans ce dernier cas, on favorisera le tracé de la densité théorique: **"curve(dnorm(x,"paramètres de la loi normale"))"**)
- **qqnorm()** pour tracer les quantiles de la distribution observées en fonction des quantiles de la distribution théorique (normale) / **qqline()** pour tracer la ligne que devrait suivre le graphique précédent si la distribution observée suivait parfaitement la distribution théorique

Lorsqu'une variable suit une loi normale, on peut alors en déduire son intervalle de confiance. **L'intervalle de confiance à 95%** d'une variable est défini par:

$$[mean(X) - 1.96sd(X); mean(X) + 1.96sd(X)]$$

Cela signifie que 95% des individus se situent dans cet intervalle. On utilise souvent les intervalles de confiance à 95% pour **déterminer si un ensemble de valeurs diffère d'un autre**, de même nature: on considère que c'est le cas si les deux intervalles de confiance sont **disjoints** (sans chevauchement).

Les relations entre deux variables quantitatives continues peuvent être décrites par différentes métriques statistiques:

- la **covariance** (fonction **“cov()”**), qui renseigne sur le niveau (et le sens: décroissant si négatif/ croissant si positif) de covariation des variables
- la **corrélation** (fonction **“cor()”**), qui renseigne sur le degré (et le sens: décroissant si négatif/ croissant si positif) d'association entre variables, c'est-à-dire la prépondérance de la relation entre variable sur les variations internes des variables
- on utilise préférentiellement la **corrélation de Pearson** (fonction **“cor()”**) quand la relation est de type **linéaire**
- on utilise préférentiellement la **corrélation de Spearman** (fonction **“cor(method =”spearman”)“**) quand la relation est **non linéaire mais monotone** (la formule donnée en cours ne s'applique que si la relation est strictement monotone)

Il est primordial de se rappeler que **le niveau de corrélation ne donne aucune certitude sur les relations de cause à effet**, seulement de potentielles indications qui'il conviendra de tester expérimentalement si on veut s'en assurer.

Il faut bien penser à étudier graphiquement le type de relation entre les deux variables avant d'utiliser les métriques de corrélation, afin de sélectionner celle la plus adaptée à notre cas. Il peut parfois être utile de **centrer-réduire** les variables pour **s'affranchir de la dimensionnalité** (sensibilité aux unités) en utilisant la fonction **“scale()”** dans R.