IUT Grand Ouest Normandie
Responsable : Alain LUCAS

Tests d'hypothèses

`Compétence 2 - Ressource 3.06`

Fiche TP $n^o 2$

Tests de comparaison de 2 moyennes

Objectif. Le but de ce TP est de vous présenter, sous R, les outils numériques permettant de réaliser un test de comparaison de 2 moyennes. On va distinguer 3 approches :

- Le test de comparaison de 2 moyennes dans le cadre de deux échantillons indépendants homoscédastiques
- Le test de comparaison de 2 moyennes dans le cadre de deux échantillons indépendants hétéroscédastiques
- Le test de comparaison de 2 moyennes dans le cadre de deux échantillons appariés

Exercice 1

Contexte. Un enseignant de statistique du Bachelor `" Science des Données"` de l'Université de Caen Normandie a réalisé deux examens sur une promotion de 34 étudiants, laquelle est composée de 21 étudiants suivant le parcours EMS (`Exploration et Modélisation Statistique`) et de 13 étudiants suivant le parcours VCOD (`Visualisation et Conception d'un Outil Décisionnel`).

L'objectif pour l'enseignant est d'essayer de prouver, pour chacun des examens, que la moyenne théorique pour les étudiants du parcours EMS (μ_{EMS}) est supérieure à la moyenne théorique pour les étudiants du parcours VCOD (μ_{VCOD}), autrement dit de mettre en œuvre le test d'hypothèses suivant :

$$\text{H}_0 : \mu_{\text{EMS}} = \mu_{\text{VCOD}} \qquad \text{versus} \qquad \text{H}_1 : \mu_{\text{EMS}} > \mu_{\text{VCOD}}$$

Les données relatives aux deux examens pour l'ensemble des étudiants de la promotion sont contenues dans un fichier nommé `Notes.cvs` et sont disponibles sur la plateforme E-Campus de l'Université de Caen Normandie

`https://ecampus.unicaen.fr`

	Id	Parcours	Note1	Note2
1	1	VCOD	9.26	9.58
2	2	VCOD	11.93	8.11
3	3	VCOD	13.38	8.71
4	4	VCOD	7.21	7.40
5	5	VCOD	12.20	8.49
6	6	VCOD	12.34	9.85
7	7	VCOD	10.40	9.62
8	8	VCOD	10.45	12.11
9	9	VCOD	10.41	8.84
10	10	VCOD	9.83	9.12

Extrait des données

Étape 1. Préliminaires

1. Créer sur votre bureau un dossier nommé `Analyse_Notes`, puis créer les sous-dossiers `Data` et `RStudio`.
2. Sous **RStudio**, créer un projet en choisissant le sous-dossier `RStudio` comme dossier de travail par défaut.
3. Télécharger le fichier de données, puis l'enregistrer dans le sous-dossier `Data`.
4. Ouvrir une fenêtre script, puis indiquer les informations suivantes en en-tête

```
#' ---
#' title: |
#'        |
#'        | \textcolor{purple}{\Huge TP Tests d'hypothèses}
#'        | \textcolor{blue}{\Large Comparaison de moyennes}
#' subtitle: |
#'           |
#'           | IUT Grand Ouest Normandie
#'           | Département - Science des Données
#'           | Campus de Lisieux
#' author: "Alain Lucas"
#' date: 20/12/2023
#' fontsize: 11pt
#' ---
```

5. Enfin, on se propose de charger un certain nombre de bibliothèques. Pour cela, écrire les instructions suivantes, puis les exécuter.

```
# ========================================================================= #
#' \begin{center} \bf{Chargement des librairies} \end{center}
# ========================================================================= #

library(dplyr)
library(ggplot2)

library(ggpubr)
library(purrr)
library(qqplotr)

library(rstatix) # t_test()
library(ggstatsplot) # ggbetweenstats()
```

Étape 2. Chargement et préparation des données

1. Écrire une instruction permettant de prendre connaissance de la structure du fichier. Vérifier que vous obtenez la sortie suivante

```
 [1] "Id\tParcours\tNote1\tNote2" "1\tVCOD\t9.26\t9.58"        "2\tVCOD\t11.93\t8.11"
 [4] "3\tVCOD\t13.38\t8.71"       "4\tVCOD\t7.21\t7.4"         "5\tVCOD\t12.2\t8.49"
 [7] "6\tVCOD\t12.34\t9.85"       "7\tVCOD\t10.4\t9.62"        "8\tVCOD\t10.45\t12.11"
[10] "9\tVCOD\t10.41\t8.84"
```

2. En déduire une instruction permettant d'importer les données dans un objet nommé `dataset` en faisant usage en particulier de l'argument `na.strings = "ABS"`. Vérifier que vous obtenez alors la structure suivante

```
'data.frame':	34 obs. of  4 variables:
 $ Id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Parcours: chr  "VCOD" "VCOD" "VCOD" "VCOD" ...
 $ Note1   : num  9.26 11.93 13.38 7.21 12.2 ...
 $ Note2   : num  9.58 8.11 8.71 7.4 8.49 ...
```

3. Il convient maintenant de définir la variable `Parcours` selon le type `factor`. Pour cela, écrire une instruction permettant d'opérer cette transformation. Via la fonction `summary()`, vérifier que vous obtenez les informations suivantes

```
       Id         Parcours      Note1           Note2
 Min.   : 1.00   EMS :13   Min.   : 7.21   Min.   : 7.31
 1st Qu.: 9.25   VCOD:21   1st Qu.:10.40   1st Qu.: 9.93
 Median :17.50             Median :11.23   Median :11.02
 Mean   :17.50             Mean   :11.44   Mean   :11.20
 3rd Qu.:25.75             3rd Qu.:12.34   3rd Qu.:12.01
 Max.   :34.00             Max.   :15.78   Max.   :17.88
                           NA's   :1
```

Étape 3. Analyse exploratoire

1. L'enseignant de statistique souhaite dans un premier temps savoir si les tests d'hypothèses ont un sens, autrement dit si effectivement la moyenne empirique de chacun des examens pour les étudiants du parcours EMS est bien supérieure à la moyenne empirique de chacun des examens pour les étudiants du parcours VCOD. Pour cela, écrire le code suivant

```
dataset %>%
  group_by(Parcours) %>%
  summarize(Mean1 = mean(Note1,
                         na.rm = TRUE),
            Std1 = sd(Note1,
                      na.rm = TRUE),
            Mean2 = mean(Note2,
                         na.rm = TRUE),
            Std2 = sd(Note2,
                      na.rm = TRUE))
```

   Exécuter, puis visualiser la sortie textuelle.

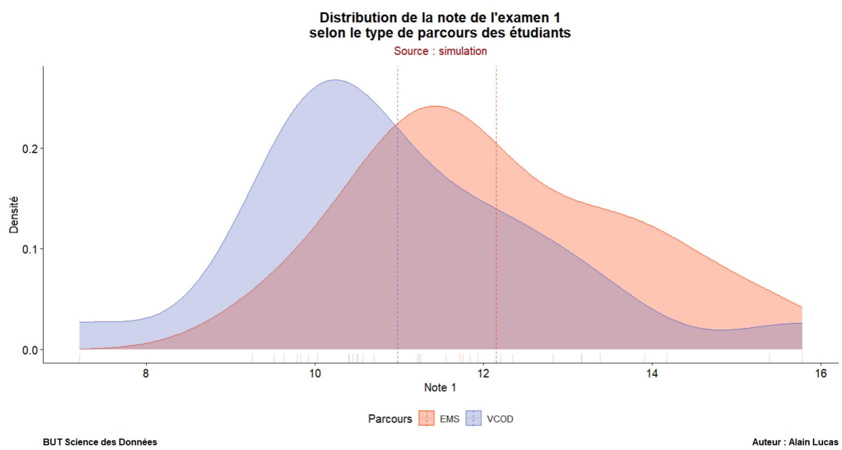
```
# A tibble: 2 × 5
  Parcours Mean1  Std1 Mean2  Std2
  <fct>    <dbl> <dbl> <dbl> <dbl>
1 EMS       12.2  1.66  11.6  2.86
2 VCOD      11.0  1.82  10.9  1.05
```

   Chacun des tests d'hypothèses a-t-il un sens ? Justifier votre réponse. Cette sortie fournit l'écart-type empirique pour chacun des échantillons et pour chacune des notes. Pour l'examen 1, est-on dans un cadre homoscédastique ou hétéroscédastique ? Justifier votre réponse. Pour l'examen 1, est-on dans un cadre homoscédastique ou hétéroscédastique ? Justifier votre réponse.

2. L'enseignant décide de se focaliser pour le moment sur la note du premier examen. Dans ce but, il décide de visualiser la répartition des notes pour chacun des groupes via une densité lissée.

```
dataset %>%
  select(Parcours,Note1) %>%
  ggdensity(x = "Note1",
            add = "mean",
            rug = TRUE,
            color = "Parcours",
            fill = "Parcours",
            alpha = 0.3,
            palette = c("#FF3D00","#5C6BC0"))+
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5,
                                  face = "bold",
                                  size = 15),
        plot.subtitle = element_text(hjust = 0.5,
                                     colour = "red4",
                                     margin = margin(b = 10)),
        plot.caption = element_text(hjust = c(0,1),
                                    face = "bold"))+
  labs(title = "Distribution de la note de l'examen 1\nselon le type de parcours des étudiants",
       subtitle = "Source : simulation",
       x = "Note 1",
       y = "Densité",
       caption = c("BUT Science des Données","Auteur : Alain Lucas"))
```

   Exécuter ces instructions, puis visualiser la représentation graphique.



   Peut-il remettre en cause l'hypothèse d'une distribution gaussienne pour chacune des distributions ? Justifier votre réponse.

3. Une alternative à cette approche subjective consiste à réaliser pour chacun des parcours un test de Shapiro-Wilk dont les hypothèses sont

$$\mathrm{H}_0 : \text{distribution gaussienne} \qquad \text{versus} \qquad \mathrm{H}_1 : \text{distribution non gaussienne}$$

Afin de mettre en œuvre le test sur chacun des groupes, écrire le code suivant

```r
# Test de Shapiro-Wilk

dataset %>%
  group_by(Parcours) %>%
  group_map(.data = .,
            .f = ~ shapiro.test(.x$Note1)) %>%
  setNames(nm = levels(dataset$Parcours))
```

Exécuter, puis visualiser la sortie textuelle.

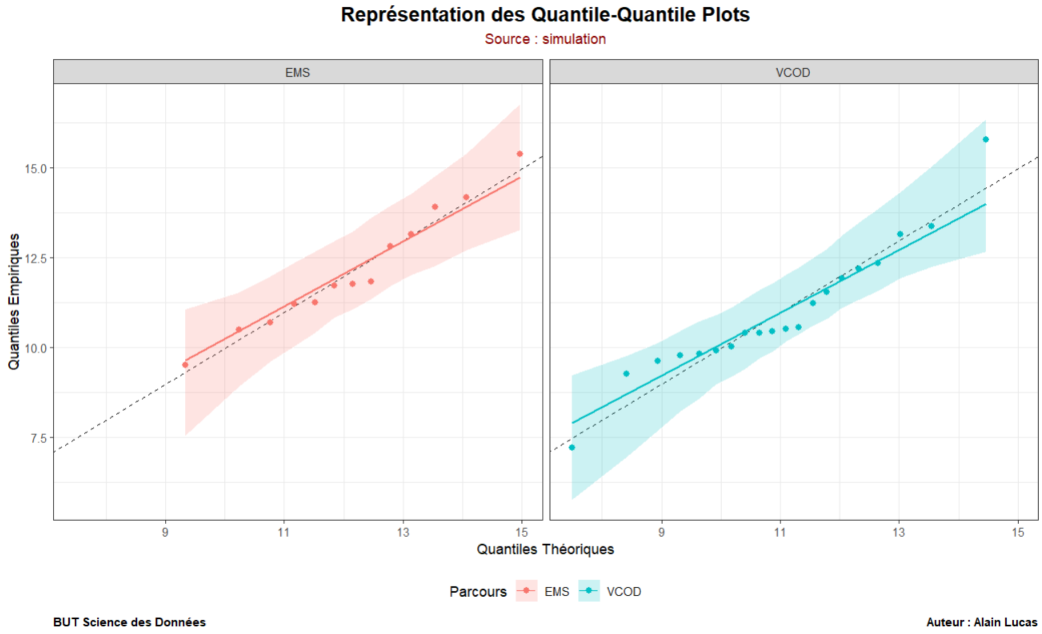
```
$EMS

	Shapiro-Wilk normality test

data:  .x$Note1
W = 0.96752, p-value = 0.863


$VCOD

	Shapiro-Wilk normality test

data:  .x$Note1
W = 0.93837, p-value = 0.2233
```

Apporter une conclusion à chacun des tests avec un niveau de signification de 5%.

4. On s'intéresse maintenant au caractère homoscédastique ou hétéroscédastique des données. Pour cela, on se propose dans un premier temps de représenter le Quantile-Quantile Plot pour les notes du premier examens pour chacun des deux groupes.

```r
dataset %>%
  ggplot(mapping = aes(colour = Parcours,
                       fill = Parcours))+
  geom_abline(slope = 1,
              intercept = 0,
              linetype = 2,
              colour = "black")+
  stat_qq_point(mapping = aes(sample = Note1),
                qtype = 4,
                size = 2)+
  stat_qq_line(mapping = aes(sample = Note1))+
  stat_qq_band(mapping = aes(sample = Note1),
               colour = "white",
               alpha = 0.2,
               bandType = "boot")+
  facet_wrap(facets = "Parcours")+
  theme_bw()+
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5,
                                  face = "bold",
                                  size = 15),
        plot.subtitle = element_text(hjust = 0.5,
                                     colour = "red4",
                                     margin = margin(b = 10)),
        plot.caption = element_text(hjust = c(0,1),
                                    face = "bold"))+
  labs(title = "Représentation des Quantile-Quantile Plots",
       subtitle = "Source : simulation",
       x = "Quantiles Théoriques",
       y = "Quantiles Empiriques",
       caption = c("BUT Science des Données","Auteur : Alain Lucas"))
```

Exécuter, puis visualiser la sortie graphique.



Que peut-on dire des deux droites des moindres carrés ? En déduire que l'on est ici plutôt dans un cadre homoscédastique.

5. Pour confirmer notre visualisation, on décide de réaliser un test de comparaison des variances de Levene dont les hypothèses sont

$$\mathrm{H}_0 : \sigma^2_{\mathrm{EMS}} = \sigma^2_{\mathrm{VCOD}} \qquad \text{versus} \qquad \mathrm{H}_1 : \sigma^2_{\mathrm{EMS}} \neq \sigma^2_{\mathrm{VCOD}}$$

Pour cela, écrire l'instruction suivante

```
dataset %>%
  levene_test(formula = Note1 ~ Parcours,
              center = median)
```

Exécuter, puis visualiser la sortie textuelle.

```
# A tibble: 1 × 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     1    31   0.00333 0.954
```

Que peut-on en déduire avec un niveau de signification de 5% ?

6. Finalement, il ressort des analyses précédentes que l'on peut supposer une distribution gaussienne pour la variable `Note1` pour chacun des groupes et que l'on est dans une situation d'homoscédasticité. En conséquence, on peut faire usage du T-test unilatéral usuel pour répondre à la problématique initiale.

   - Une première solution consiste à faire usage de la fonction de base `t.test()`

```
dataset %>%
  t.test(data = .,
         Note1 ~ Parcours,
         alternative = "greater",
         paired = FALSE,
         var.equal = TRUE)
```

Exécuter, puis visualiser la sortie textuelle.

```
        Two Sample t-test

data:  Note1 by Parcours
t = 1.8665, df = 31, p-value = 0.03573
alternative hypothesis: true difference in means between group EMS and group VCOD is greater than 0
95 percent confidence interval:
 0.1073197       Inf
sample estimates:
 mean in group EMS mean in group VCOD
          12.15077           10.97900
```

Que peut-on en déduire avec un niveau de signification de 5% ? Justifier votre réponse.

- Une alternative consiste à faire usage de la fonction `t_test()` de la bibliothèque `rstatix`

```
dataset %>%
  t_test(data = .,
         formula = Note1 ~ Parcours,
         alternative = "greater",
         paired = FALSE,
         var.equal = TRUE)
```

Exécuter, puis visualiser la sortie textuelle.

```
# A tibble: 1 x 8
  .y.   group1 group2    n1    n2 statistic    df      p
* <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>  <dbl>
1 Note1 EMS    VCOD      13    20      1.87    31 0.0357
```

Vérifier que l'on obtient un résultat en tout point identique malgré une présentation un peu différente.

- Une seconde alternative consiste à faire usage de la fonction `compare_means()` de la bibliothèque `ggpubr`

```
dataset %>%
  compare_means(formula = Note1 ~ Parcours,
                data = .,
                method = "t.test",
                paired = FALSE,
                var.equal = TRUE,
                p.adjust.method = "none",
                alternative = "less") %>%
  select(-c(p.adj,p.format,p.signif)) %>%
  mutate(p.value = round(100*p,2)) %>%
  select(c(method,.y.,group1,group2,p.value,-p))
```

Exécuter, puis visualiser la sortie textuelle.

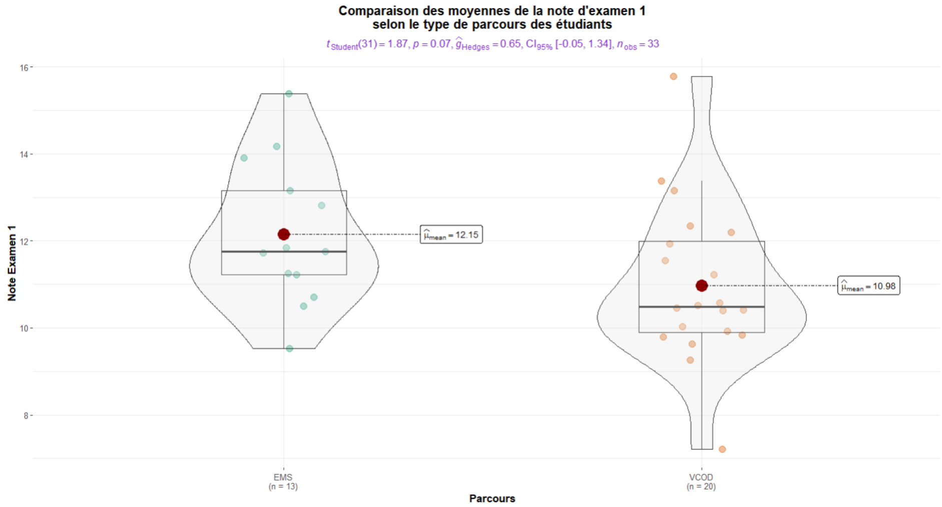
```
# A tibble: 1 x 5
  method .y.   group1 group2 p.value
  <chr>  <chr> <chr>  <chr>    <dbl>
1 T-test Note1 EMS    VCOD      3.57
```

Attention, il faut noter ici que pour obtenir le même résultat, il faut utiliser l'argument `alternative = "less"`, i.e. que le test est réalisé à l'envers des tests précédents !!!

7. On peut également proposer des représentations graphiques pour afficher les résultats. On pourra se tourner par exemple vers la fonction `ggbetweenstats()` de la bibliothèque `ggstatsplot`.

```
dataset %>%
  ggbetweenstats(x = Parcours,
                 y = Note1,
                 type = "parametric",
                 p.adjust.method = "none",
                 alternative = "less",
                 var.equal = TRUE,
                 results.subtitle = TRUE,
                 violin.args = list(width = 0.5,
                                    alpha = 0.2,
                                    fill ="grey85",
                                    na.rm = TRUE)) +
  theme(plot.caption = element_blank(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5,
                                     colour = "purple3"))+
  labs(title = "Comparaison des moyennes de la note d'examen 1\nselon le type de parcours des étudiants",
       x = "Parcours",
       y = "Note Examen 1")
```

Exécuter, puis visualiser la sortie graphique.



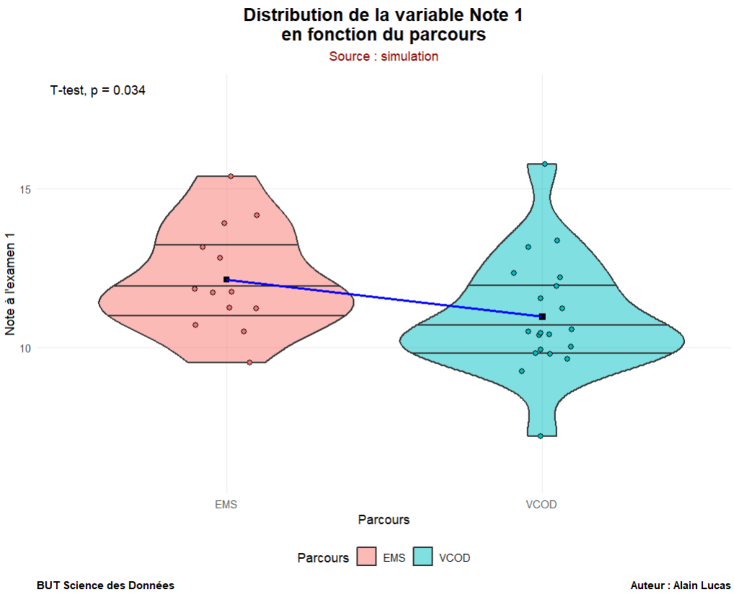
La principale problématique est que la p-value est obtenue pour un test bilatéral et non un test unilatéral. En pratique, il faut diviser cette valeur par 2 pour obtenir la véritable p-value.

Une alternative consiste à faire usage de la fonction `stat_compare_means()` de la bibliothèque `ggpubr`

```
dataset %>%
  ggplot(mapping = aes(x = Parcours,
                       y = Note1))+
  geom_violin(mapping = aes(fill = Parcours),
              alpha = 0.5,
              linetype = 1,
              linewidth = 0.7,
              bw = 0.7,
              draw_quantiles = c(0.25,0.5,0.75))+
  geom_jitter(mapping = aes(fill = Parcours),
              shape = 21,
              size = 2,
              width = 0.1,
              show.legend = FALSE)+
  stat_summary(geom = "point",
               fun = "mean",
               size = 2,
               shape = 22,
               colour = "black",
               fill = "black",
               show.legend = FALSE)+
  stat_summary(geom = "line",
               fun = "mean",
               group = 1,
               linewidth = 1,
               colour = "blue",
               show.legend = FALSE)+
```

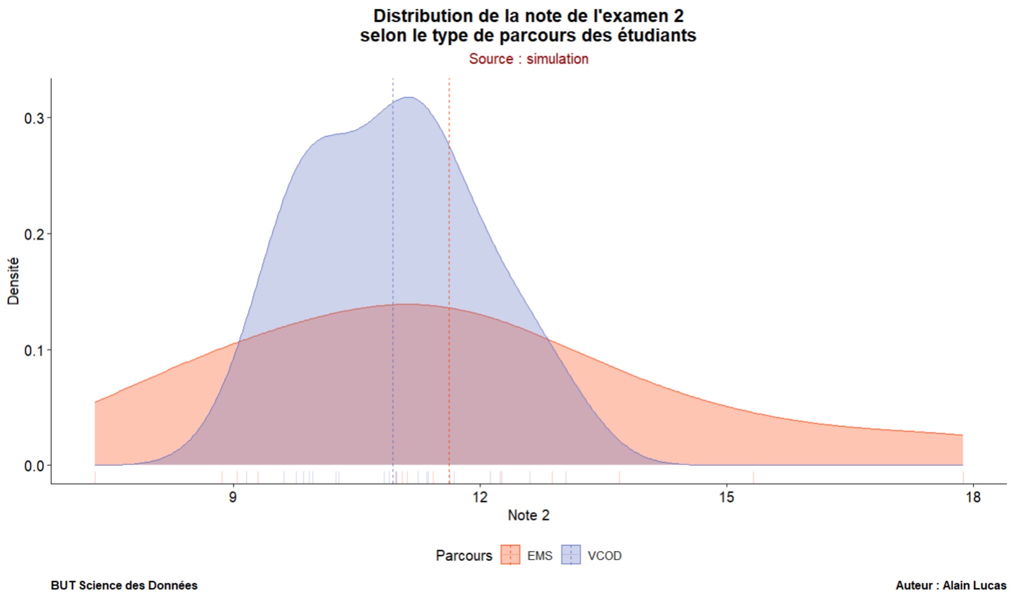
```
stat_compare_means(method = "t.test",
                   paired = FALSE,
                   geom = "text",
                   label.x = 0.5,
                   label.y = 18,
                   method.args = list(alternative = "greater",
                                      var.equal = TRUE))+
scale_y_continuous(limits = c(6,18))+
theme_minimal()+
theme(legend.position = "bottom",
      plot.title = element_text(hjust = 0.5,
                                face = "bold",
                                size = 15),
      plot.subtitle = element_text(hjust = 0.5,
                                   colour = "red4",
                                   margin = margin(b = 10)),
      plot.caption = element_text(hjust = c(0,1),
                                  face = "bold"),
      panel.grid.minor.y = element_blank())+
labs(title = "Distribution de la variable Note 1\nen fonction du parcours",
     subtitle = "Source : simulation",
     x = "Parcours",
     y = "Note à l'examen 1",
     caption = c("BUT Science des Données","Auteur : Alain Lucas"))
```

Exécuter, puis visualiser la sortie graphique.



Dans les deux cas, on peut noter que l'on est amené à rejeter l'hypothèse nulle au profit de l'hypothèse alternative et donc à déduire de cette analyse statistique que la moyenne théorique pour le groupe EMS est significativement supérieure à la moyenne théorique pour le groupe VCOD sur le premier examen.

8. L'enseignant se propose maintenant de réaliser un travail similaire pour la note du deuxième examen dont il a pu observer que la moyenne empirique pour le groupe EMS étati bien supérieure à la moyenne empirique pour le groupe VCOD. Il décide dans un premier temps de visualiser la distribution de la note de ce second examen pour chacun des groupes. Vérifier que vous obtenez le graphique suivant



Peut-on remettre en cause l'hypothèse d'une distribution gaussienne dans l'un et l'autre cas ? Justifier votre réponse. Est-on plutôt dans un cadre homoscédastique ou hétéroscédastique ? Justifier votre réponse.

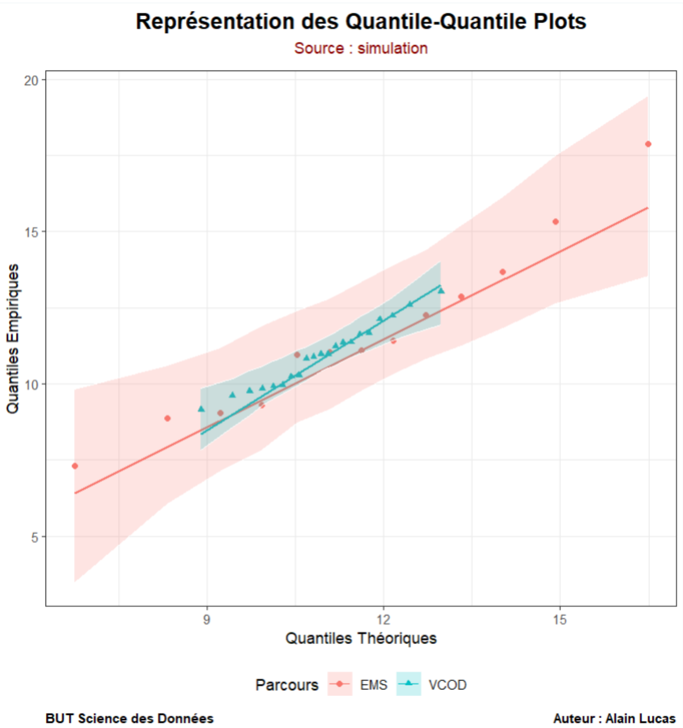
9. Pour avoir une réponse objective à chacune des précédentes questions, l'enseignant décide de réaliser un test de Shapiro-Wilk et un test d'égalité des variances. Vérifier que vous obtenez les sorties suivantes :

```
$EMS

        Shapiro-Wilk normality test

data:  .x$Note2
W = 0.95717, p-value = 0.7096


$VCOD

        Shapiro-Wilk normality test

data:  .x$Note2
W = 0.97277, p-value = 0.7934
```

```
# A tibble: 1 × 4
    df1   df2 statistic      p
  <int> <int>     <dbl>  <dbl>
1     1    32      7.25 0.0112
```

Que pouvez-vous en conclure pour l'un et l'autre des tests d'hypothèses pour un niveau de signification de 5% ? Justifier vos réponses.

10. Pour compléter son analyse, l'enseignant a décidé de représenter le Quantile-Quantile Plot pour chacun des groupes.



Observer que le points sont plutôt distribués le long d'une droite témoignant d'une forte compatibilité avec l'hypothèse gaussienne et que par ailleurs les droites ne sont manifestement pas parallèles témoignant d'un cadre plutôt hétéroscédastique.

11. L'enseignant décide maintenant de mettre en øeuvre le test paramétrique de comparaison des moyennes selon la méthode de Welsh du fait d'un cadre hétéroscédastique. Vérifier que vous obtenez les sorties suivantes :

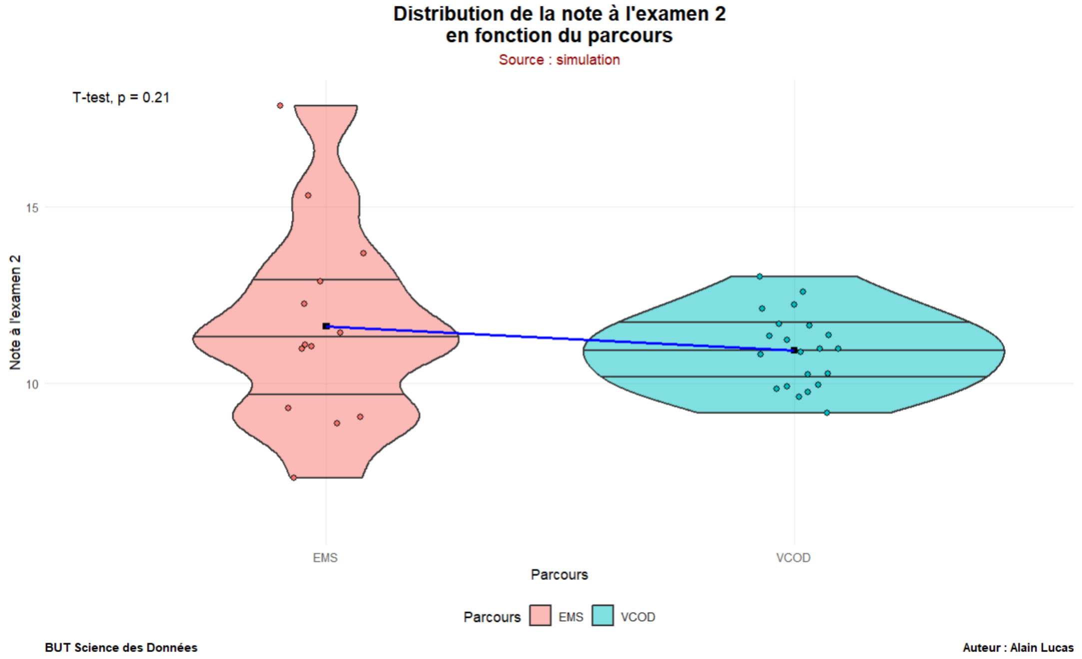
```
        Welch Two Sample t-test

data:  Note2 by Parcours
t = 0.82662, df = 14.033, p-value = 0.2111
alternative hypothesis: true difference in means between group EMS and group VCOD is greater than 0
95 percent confidence interval:
 -0.7726678        Inf
sample estimates:
 mean in group EMS mean in group VCOD
         11.62308          10.93952
```

```
# A tibble: 1 × 8
  .y.   group1 group2    n1    n2 statistic    df     p
* <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl>
1 Note2 EMS    VCOD      13    21     0.827  14.0 0.211
```

```
# A tibble: 1 × 5
  method .y.   group1 group2 p.value
  <chr>  <chr> <chr>  <chr>    <dbl>
1 T-test Note2 EMS    VCOD      21.1
```

Que doit-il décider dans le cas présent pour un niveau de signification de 5% ? Justifier votre réponse.

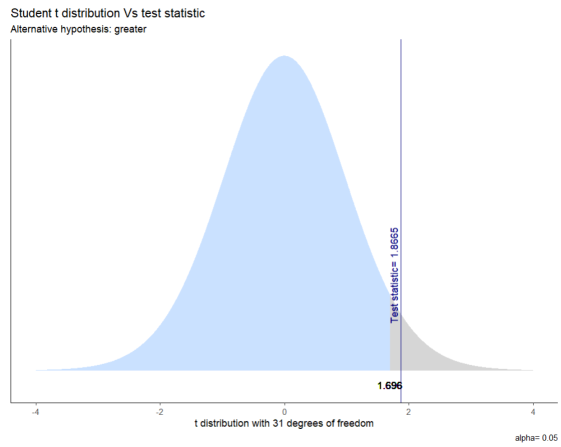
12. Enfin, l'enseignant souhaite visualiser les distributions et dans le même temps le résultat du test. Vérifier que vous obtenez la sortie suivante

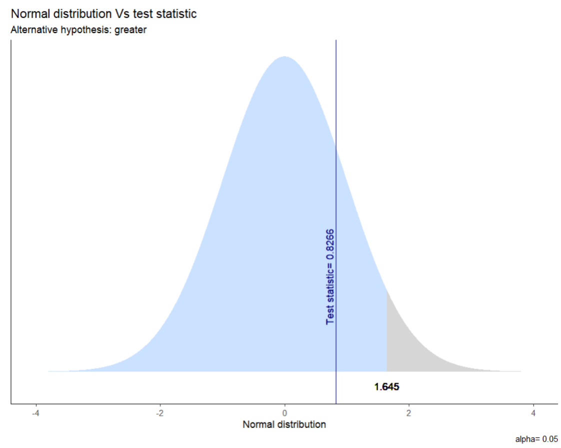


13. Une alternative consiste également à faire usage de la fonction `ggttest()` de la bibliothèque `gginference`

```
library(gginference)


t.test(dataset$Note1 ~ dataset$Parcours,
       alternative = "greater",
       paired = FALSE,
       var.equal = TRUE) -> result1

ggttest(result1)

t.test(dataset$Note2 ~ dataset$Parcours,
       alternative = "greater",
       paired = FALSE,
       var.equal = FALSE) -> result2

ggttest(result2)
```

Exécuter, puis vérifier que vous obtenez les sorties graphiques suivantes





A gauche, on est amené à rejeter l'hypothèse nulle car la valeur de la statistique de test est plus grande que la valeur critique. A contrario, à droite, on est amené à conserver l'hypothèse nulle car la statistique de test est plus faible que la valeur critique.

Exercice 2

Contexte. Un laboratoire pharmaceutique a développé une nouvelle molécule dont l'objectif est de faire perdre du poids chez les patients en situation d'obésité. Afin de vérifier l'intérêt de cette molécule, un biologiste a pour mission de la tester sur un échantillon de 25 souris considérées en surpoids.

A l'issu de cette expérimentation, le biologiste a collecté les données comportant le poids avant (`Weight_before`) et après (`Weight_after`) administration du traitement. Ces données sont contenues dans un fichier nommé `Experiment.csv` disponible sur la plateforme E-Campus de l'Université de Caen Normandie

https://ecampus.unicaen.fr

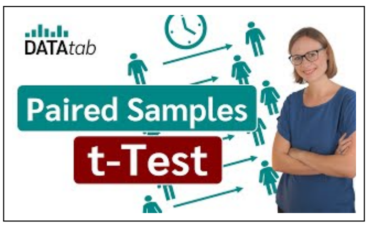
	Weight_before	Weight_after
1	175.8587	106.55385
2	205.5486	167.24267
3	221.6888	119.29033
4	153.0860	149.54585
5	208.5825	121.92154
6	210.1211	183.06893
7	188.5052	135.73221
8	189.0674	128.71680
9	188.7110	134.96226
10	182.1992	101.12720
11	190.4561	114.97142
12	180.0323	84.59881
13	184.4749	109.77020
14	201.2892	141.17118
15	219.1899	136.02307
16	197.7943	193.48489

Extrait des données

La problématique pour ce biologiste est très simple : peut-il décider au vu de son expérimentation, avec un niveau de signification de 5%, que le traitement est significativement efficace pour faire baisser le poids de souris en surpoids ? Ici, le biologiste est dans une situation spécifique car les variables `Weight_before` et `Weight_after` sont observées sur des unités statistiques identiques : on parle d'échantillon apparié ou `paired sample`. Le test statistique pour répondre à la problématique est donc le `T-Test` mais pour données appariées ou `Paired Sample T-Test`. Il se décrit selon le contexte de la manière suivante :

$$\mathrm{H}_0 : \mu_{\text{before}} = \mu_{\text{after}} \qquad \text{versus} \qquad \mathrm{H}_1 : \mu_{\text{before}} > \mu_{\text{after}}$$

On est ici dans le cadre d'un test unilatéral supérieur dont l'objectif consiste à prouver que le poids moyen théorique avant traitement est inférieur au poids moyen théorique après traitement.



1. Accéder à la plateforme E-Campus de l'Université de Caen Normandie

   https://ecampus.unicaen.fr

   puis télécharger le fichier de données `Experiment.csv`.
2. Créer sur votre espace de travail un dossier nommé `Experiment`, puis deux sous-dossiers respectivement nommé `Data` et `RStudio`.
3. Sans le logiciel **RStudio**, créer un projet en choisissant le sous-dossier `RStudio` comme dossier de travail par défaut. Enregistrer par ailleurs le fichier de données dans le sous-dossier `Data`.
4. Ouvrir un fichier script (`Ctrl+Shift+N`), puis indiquer par exemple l'en-tête suivant :

```
#' ---
#' title: |
#'        |
#'        | \textcolor{purple}{\Huge TP Tests d'hypothèses}
#'        | \textcolor{blue}{\Large Comparaison de moyennes - Cas de données appariées}
#' subtitle: |
#'           |
#'           | IUT Grand Ouest Normandie
#'           | Département - Science des Données
#'           | Campus de Lisieux
#' author: "Alain Lucas"
#' date: 21/12/2023
#' fontsize: 11pt
#' ---
```

5. Dans un second temps, indiquer les bibliothèques suivantes (on s'assurera de leur présence sur le disque en amont) :

```
# ========================================================================== #
#' \begin{center} \bf{Chargement des librairies} \end{center}
# ========================================================================== #

library(ggplot2)
library(dplyr)
library(tidyr)    # pivot_longer()
library(stringr)

library(purrr)    # map()
library(rstatix) # shapiro_test()

library(tibble) # num()
```

   Exécuter, puis vérifier le succès de l'opération.
6. Écrire une instruction permettant de visualiser les 10 premières lignes du fichier `Experiment.csv`. Déterminer alors les caractéristiques du fichier.

```
 [1] "Weight_before\tWeight_after" "175.86\t106.55"            "205.55\t167.24"
 [4] "221.69\t119.29"              "153.09\t149.55"            "208.58\t121.92"
 [7] "210.12\t183.07"              "188.51\t135.73"            "189.07\t128.72"
[10] "188.71\t134.96"
```

   En déduire une instruction permettant de charger ces données dans un objet nommé `dataset`. Exécuter, puis vérifier le succès de l'opération.

```
'data.frame':   25 obs. of  2 variables:
 $ Weight_before: num  176 206 222 153 209 ...
 $ Weight_after : num  107 167 119 150 122 ...
```

7. En l'état, le jeu de données n'est pas prêt pour une analyse statistique. Il convient de créer une variable `Treatment` comprenant deux modalités ordonnées : `before` et `after` ; et une variable `Weight` contenant le poids des souris avant et après l'administration du traitement. Pour cela, on va effectuer un pivot via la fonction `pivot_longer` de la bibliothèque `tidyr` ainsi qu'une définition convenable du type de la variable `Treatment`

```
# ========================================================================== #
#' \begin{center} \bf{Préparation des données} \end{center}
# ========================================================================== #


dataset |>
  pivot_longer(cols = c(Weight_before,Weight_after),
               names_to = "Treatment",
               values_to = "Weight",
               names_pattern = "_(.*)") |>
  mutate(Treatment = factor(Treatment,
                            ordered = TRUE,
                            levels = c("before","after"))) -> data

str(data)
```

Exécuter ce code, puis vérifier le succès de l'opération.

```
tibble [50 × 2] (S3: tbl_df/tbl/data.frame)
 $ Treatment: Ord.factor w/ 2 levels "before"<"after": 1 2 1 2 1 2 1 2 1 2 ...
 $ Weight   : num [1:50] 176 107 206 167 222 ...
```

8. Maintenant que les données sont prêtes, il s'agit de s'assurer que le test de comparaison présente un intérêt. En d'autres termes, il faut s'assurer que la moyenne empirique du poids après traitement est bien inférieur à la moyenne empirique du poids avant traitement. Pour cela, écrire les instructions suivantes

```
# ========================================================================= #
#' \begin{center} \bf{Analyse exploratoire des données} \end{center}
# ========================================================================= #

data |>
  group_by(Treatment) |>
  summarize(Count = n(),
            Mean = mean(Weight,na.rm = TRUE),
            sd = sd(Weight,na.rm = TRUE))
```

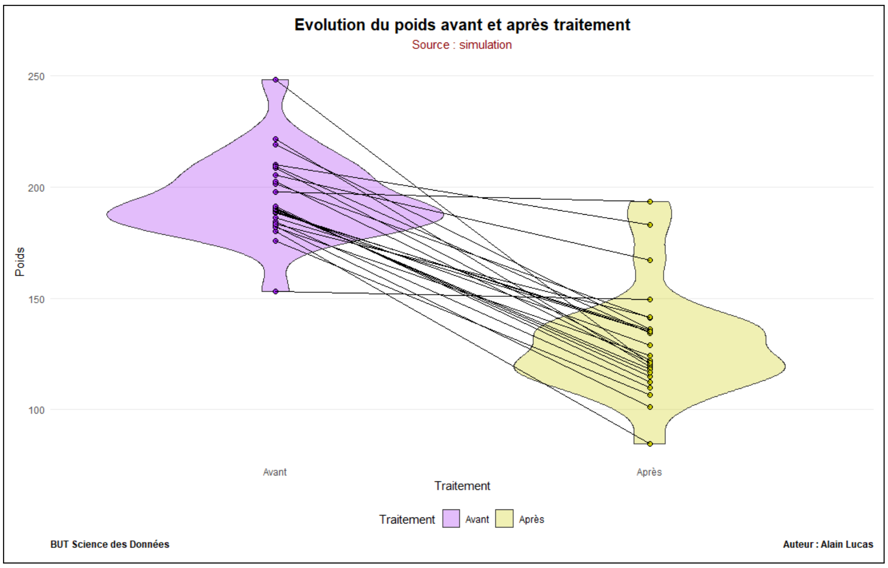
Exécuter, puis visualiser la sortie textuelle. Le test d'hypothèses a-t-il un intérêt dans le cas présent ? Justifier votre réponse.

```
# A tibble: 2 × 4
  Treatment Count  Mean    sd
  <ord>     <int> <dbl> <dbl>
1 before       25  195.  18.5
2 after        25  130.  24.2
```

On peut également proposer une représentation graphique pour rendre compte de l'intérêt du test. On pourra par exemple écrire les instructions suivantes

```
data |>
  ggplot(mapping = aes(x = Treatment,
                       y = Weight))+
  geom_violin(mapping = aes(fill = Treatment),
              alpha = 0.3)+
  geom_point(mapping = aes(fill = Treatment),
             shape = 21,
             size = 2,
             colour = "black",
             show.legend = FALSE)+
  scale_x_discrete(labels = c("Avant","Après"))+
  scale_fill_manual(name = "Traitement",
                    values = c("purple","yellow3"),
                    labels = c("Avant","Après"))+
  geom_segment(data = data |>
                 pivot_wider(names_from = Treatment,
                             values_from = Weight,
                             values_fn = list) |>
                 unnest(cols = c("before","after")),
               mapping = aes(x = 1,
                             xend = 2,
                             y = before,
                             yend = after))+
  theme_minimal()+
  theme(legend.position = "bottom",
        panel.grid.major.x = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5,
                                  face = "bold",
                                  size = 15),
        plot.subtitle = element_text(hjust = 0.5,
                                     colour = "red4",
                                     margin = margin(b = 10)),
        plot.caption = element_text(hjust = c(0,1),
                                    face = "bold"))+
  labs(title = "Evolution du poids avant et après traitement",
       subtitle = "Source : simulation",
       x = "Traitement",
       y = "Poids",
       caption = c("BUT Science des Données","Auteur : Alain Lucas"))
```

Exécuter, puis visualiser la sortie graphique.



9. Pour mettre en œuvre le test d'hypothèses, il convient maintenant de s'assurer que les hypothèses sont valides, en particulier que l'écart entre les deux variables est en accord avec une distribution gaussienne. Pour cela, on va reconstituer les données sur deux colonnes : `before` et `after` ; puis calculer l'écart entre les deux variables pour chacune des souris : création de la variable `diff` représentant donc la perte de poids.

```
# ======================================================================== #
#' \begin{center} \bf{Test d'hypothèses} \end{center}
# ======================================================================== #

#' Contrôle de l'hypothèse gaussienne des écarts

data |>
  pivot_wider(names_from = Treatment,
              values_from = Weight,
              values_fn = list) |>
  unnest(cols = c("before","after")) |>
  mutate(diff = after - before) -> global.data
```

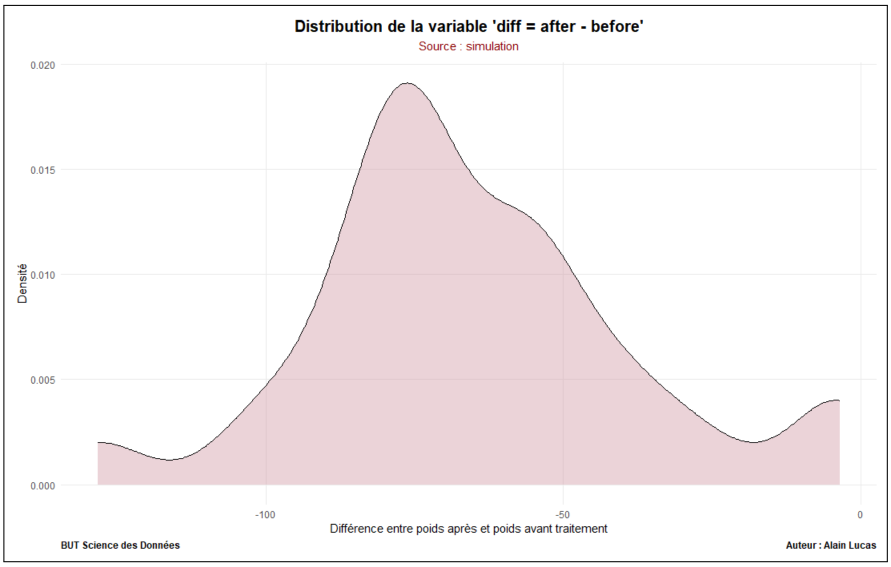
Exécuter, puis vérifier le succès de l'opération.

```
tibble [25 × 3] (S3: tbl_df/tbl/data.frame)
 $ before: num [1:25] 176 206 222 153 209 ...
 $ after : num [1:25] 107 167 119 150 122 ...
 $ diff  : num [1:25] -69.31 -38.31 -102.4 -3.54 -86.66 ...
```

Maintenant,on se propose de visualiser la distribution de cette variable `diff`.

```
global.data |>
  ggplot(mapping = aes(x = diff))+
  geom_density(fill = "pink3",
               alpha = 0.4,
               bw = 8)+
  theme_minimal()+
  theme(panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5,
                                  face = "bold",
                                  size = 15),
        plot.subtitle = element_text(hjust = 0.5,
                                     colour = "red4",
                                     margin = margin(b = 10)),
        plot.caption = element_text(hjust = c(0,1),
                                    face = "bold"))+
  labs(title = "Distribution de la variable 'diff = after - before'",
       subtitle = "Source : simulation",
       x = "Différence entre poids après et poids avant traitement",
       y = "Densité",
       caption = c("BUT Science des Données","Auteur : Alain Lucas"))
```

Exécuter, puis visualiser la représentation graphique. Que représente la courbe sur le graphique ?



Peut-on remettre en cause l'hypothèse d'une distribution gaussienne dans le cas présent ? Justifier votre réponse.

On se propose de compléter cette approche subjective par un test de Shapiro-Wilk dont les hypothèses sont

$$\mathrm{H}_0 : \text{distribution gaussienne} \qquad \text{versus} \qquad \mathrm{H}_1 : \text{distribution non gaussienne}$$

Pour cela, écrire les instructions suivantes

```
#' Test de Shapiro-Wilk

global.data |>
  select(diff) |>
  map(.f = ~ .x |> shapiro.test())


global.data |>
  shapiro_test(diff)


global.data |>
  select(diff) |>
  map(.f = ~ .x |>
        shapiro_test() |>
        mutate(variable = NULL)) |>
  bind_rows(.id = "Variable")
```

Exécuter, puis vérifier que l'on obtient dans les trois approches le même résultat.

```
> global.data |>
+   select(diff) |>
+   map(.f = ~ .x |> shapiro.test())
$diff

	Shapiro-Wilk normality test

data:  .x
W = 0.95826, p-value = 0.3809


> global.data |>
+   shapiro_test(diff)
# A tibble: 1 × 3
  variable statistic     p
  <chr>        <dbl> <dbl>
1 diff         0.958 0.381
> global.data |>
+   select(diff) |>
+   map(.f = ~ .x |>
+         shapiro_test() |>
+         mutate(variable = NULL)) |>
+   bind_rows(.id = "Variable")
# A tibble: 1 × 3
  Variable statistic p.value
  <chr>        <dbl>   <dbl>
1 diff         0.958   0.381
```

Que doit-on décider avec un niveau de signification de 5% ? Justifier votre réponse.

10. Maintenant que l'on a vérifié l'hypothèse gaussienne, il s'agit de réaliser le test. Pour cela, on peut faire usage de la fonction de base `t.test()` ou de la fonction `t_test()` de la bibliothèque `rstatix`.

```r
#' Test de comparaison de moyennes - paired = TRUE

t.test(formula = data$Weight ~ data$Treatment,
       alternative = "greater",
       paired = TRUE)

data |>
  t_test(formula = Weight ~ Treatment,
         alternative = "greater",
         mu = 0,
         paired = TRUE,
         detailed = TRUE) |>
  mutate(statistic = num(statistic,digits = 3))
```

Exécuter, puis vérifier que l'on obtient un résultat identique dans les deux cas.

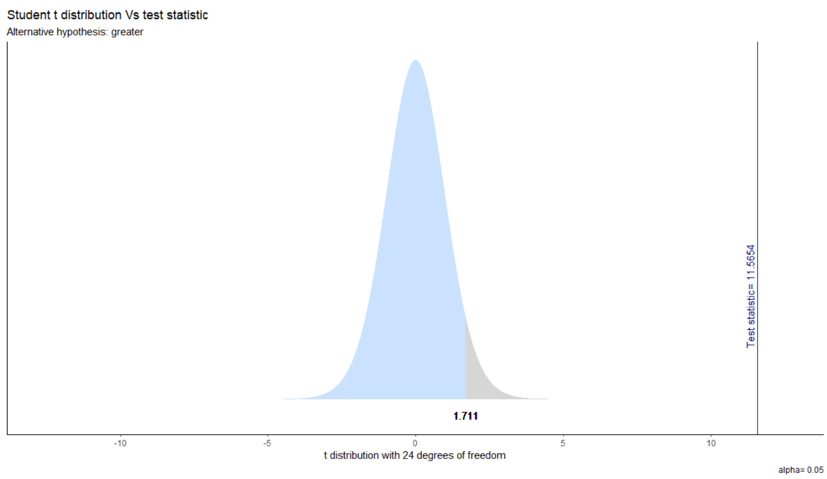
```
> t.test(formula = data$Weight ~ data$Treatment,
+        alternative = "greater",
+        paired = TRUE)

	Paired t-test

data:  data$Weight by data$Treatment
t = 11.565, df = 24, p-value = 1.336e-11
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 55.46594      Inf
sample estimates:
mean difference
        65.0956

> data |>
+   t_test(formula = Weight ~ Treatment,
+          alternative = "greater",
+          mu = 0,
+          paired = TRUE,
+          detailed = TRUE) |>
+   mutate(statistic = num(statistic,digits = 3))
# A tibble: 1 × 13
  estimate .y.    group1 group2    n1    n2 statistic        p    df conf.low conf.high method alternative
     <dbl> <chr>  <chr>  <chr>  <int> <int> <num:.3!>    <dbl> <dbl>    <dbl>     <dbl> <chr>  <chr>
1     65.1 Weight before after     25    25    11.565 1.34e-11    24     55.5       Inf T-test greater
```

Finalement, que doit en conclure le biologiste avec un niveau de signification de 5% ? Justifier votre réponse.

11. Une dernière approche consiste à prendre une décision selon la valeur critique. Pour cela, on va faire usage de la fonction `ggttest()` de la bibliothèque `gginference`.

```r
#' Utilisation de la librairie gginference

library(gginference)

t.test(formula = data$Weight ~ data$Treatment,
       alternative = "greater",
       paired = TRUE)  -> test

ggttest(test)
```

Exécuter, puis visualiser la sortie graphique.



Observer que la valeur de la statistique de test est très nettement au delà de la valeur critique de 1.711, amenant au rejet de l'hypothèse nulle avec une confiance importante.